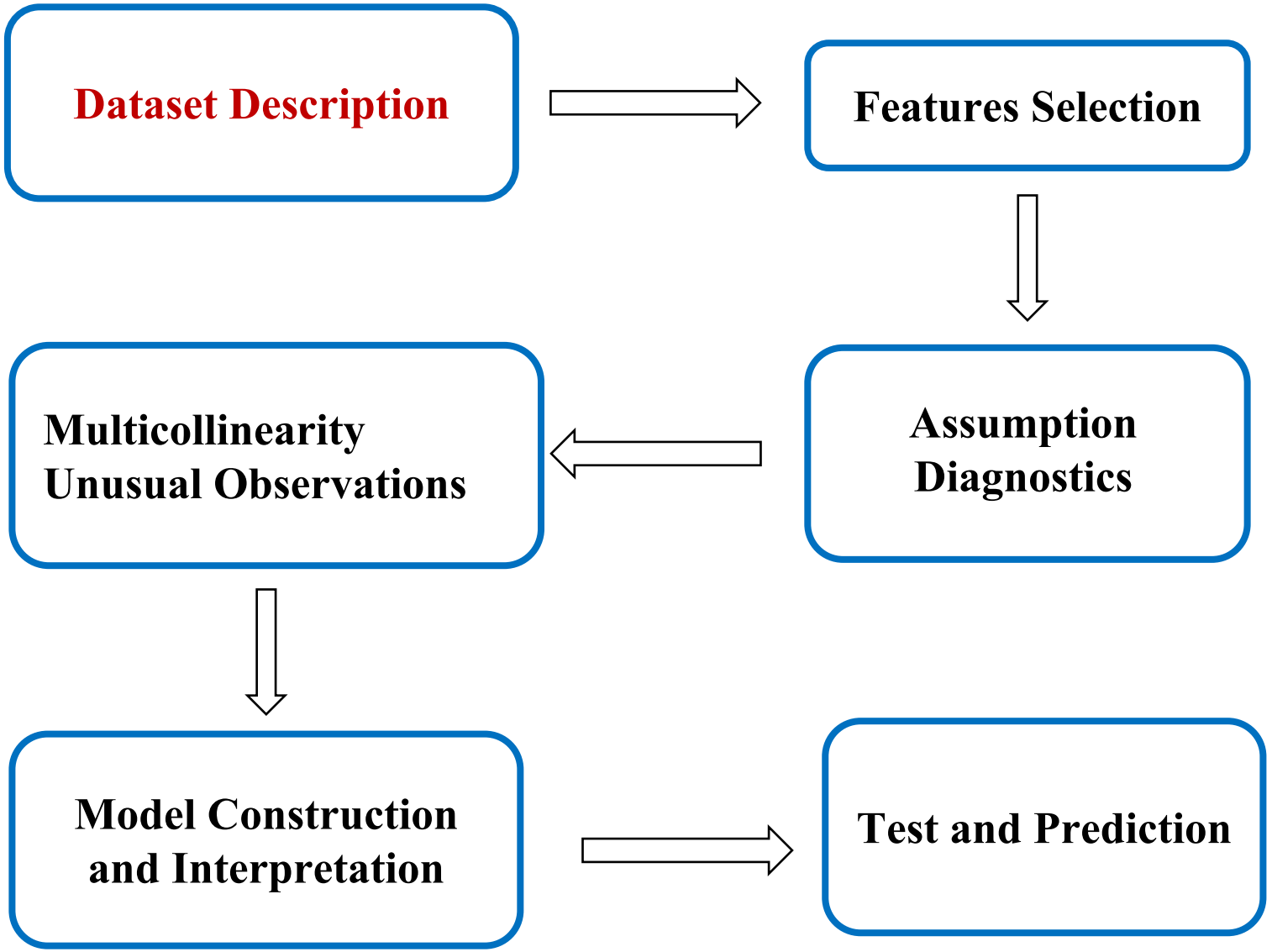
Abalone Age Prediction

Yunhao Li
Siheng Huang
Jiechun Lin

ABSTRACT

- Dataset Description
- Feature Selection
- Assumption Diagnostics
- Multicollinearity
- Unusual Observations
- Model Construction and Interpretation
- Test and Prediction


Dataset Description
Features Selection
Assumption Diagnostics
Multicollinearity Unusual Observations
Model Construction and Interpretation
Test and Prediction

Introduction and Goal

- Abalone is a desirable food yet usually expensive. The economic value of abalone is positively correlated with age; however, the technical means of testing abalone age are often relatively expensive and time-consuming.
- Therefore, a model that only needs external data of abalone to predict the age is very practical. We hope we can establish a model to predict the age of abalone though linear regression in this project.

Data Description

Variable Name	Description	Data Type	Measurement unit
Sex	M, F, and I (infant)	Nominal	NA
Length	Longest shell measurement	Continuous	mm
Diameter	perpendicular to length	Continuous	mm
Height	with meat in shell	Continuous	mm
Whole weight	whole abalone	Continuous	grams
Shucked weight	weight of meat	Continuous	grams
Viscera weight	gut weight (after bleeding)	Continuous	grams
Shell weight	after being dried	Continuous	grams
Rings	+1.5 gives the age in years	Integer	NA

- This dataset includes information on more than 4000 abalones
- In general, we will use rings as the response and other variables as independent variables to build a regression model.

The original dataset can be acessed at https://archive.ics.uci.edu/ml/datasets/abalone.

Data Exploration

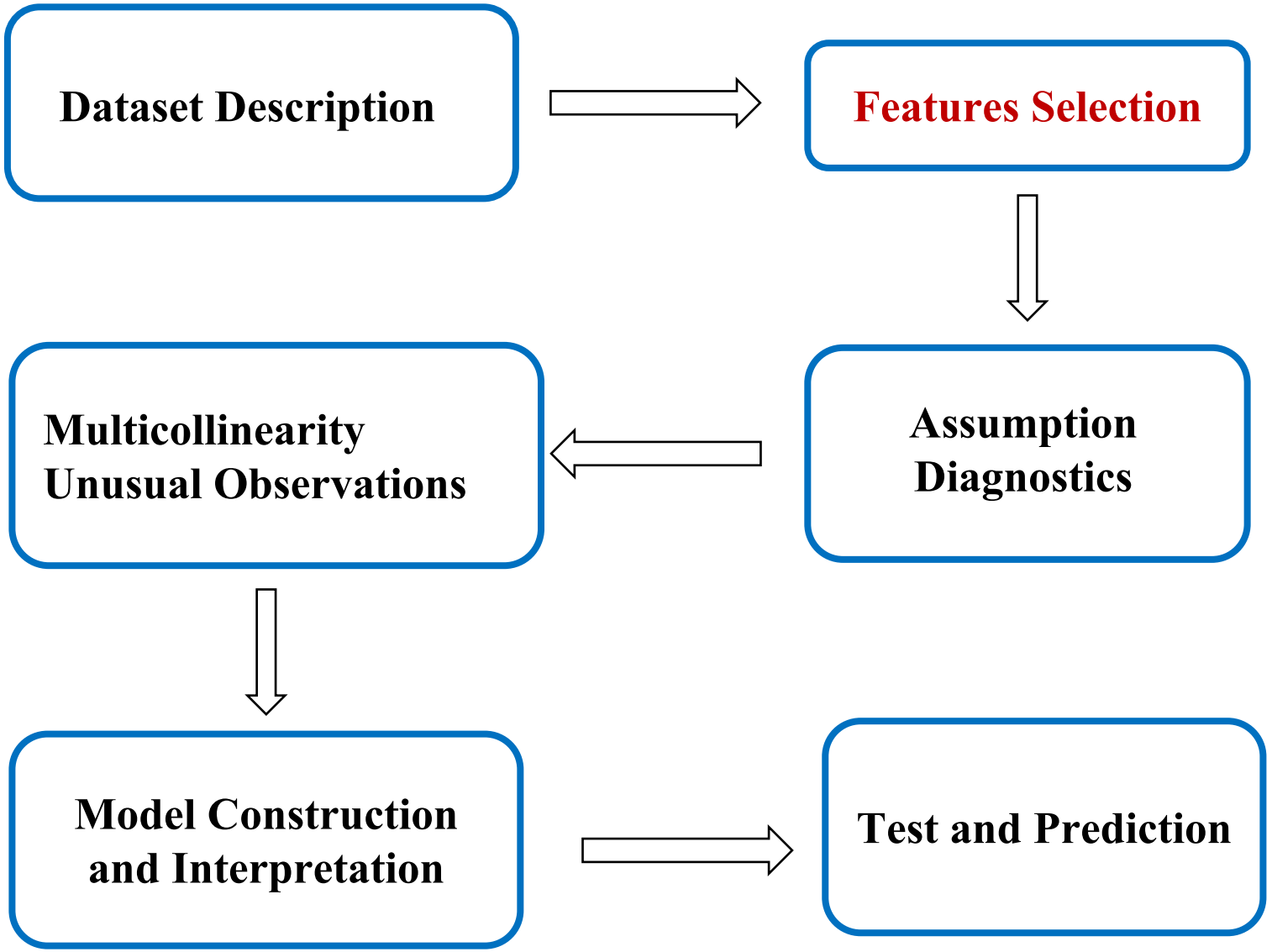
The independent Variable (also called features) can be divided into 3 groups:

Independent Variables

- **Gender** — **Discrete data**
- **Length** — **Continuous Data**
- **Weight** — **Continuous Data**

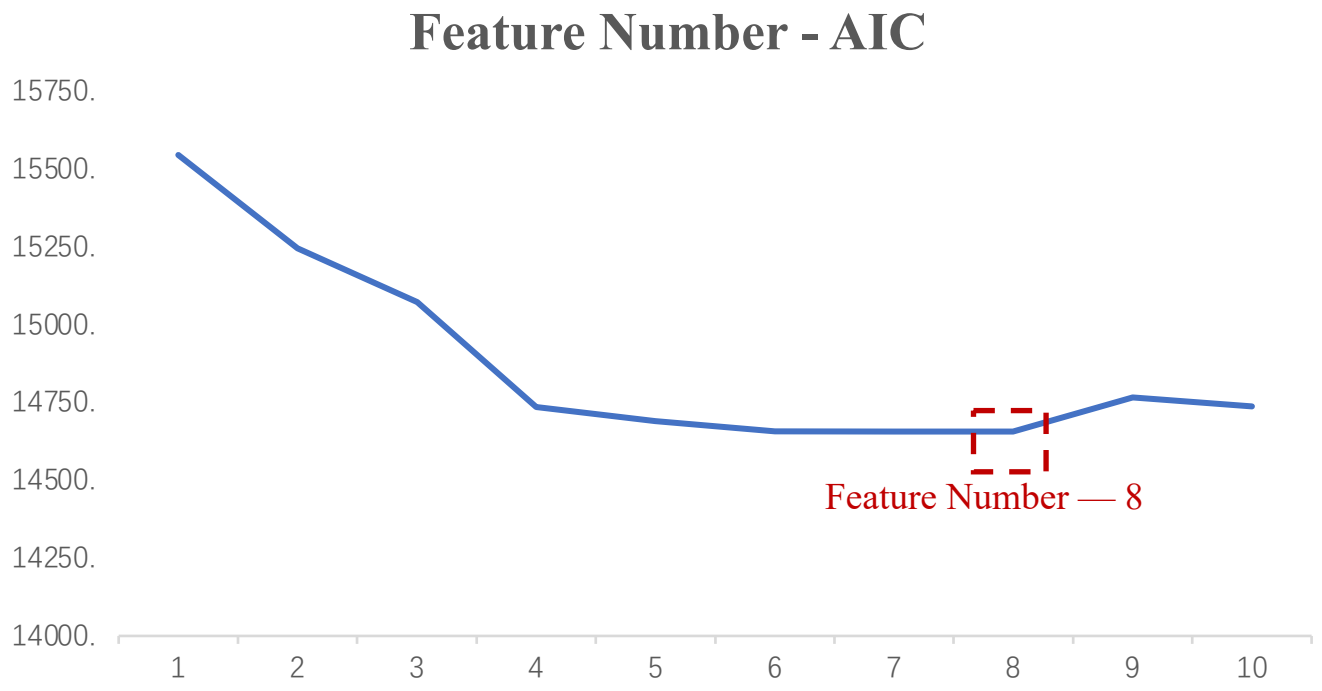
Gender: M, F, I
Apply one-hot encode

Gender	Gender_M	Gender_F	Gender_I
M	1	0	0
F	0	1	0
I	0	0	1


Dataset Description
Features Selection
Assumption Diagnostics
Multicollinearity Unusual Observations
Model Construction and Interpretation
Test and Prediction

Feature Selection

Implement **Forward Search** to the features and choose **AIC** as the criterion.



- We chose AIC because we aim to select a good **model for prediction.**

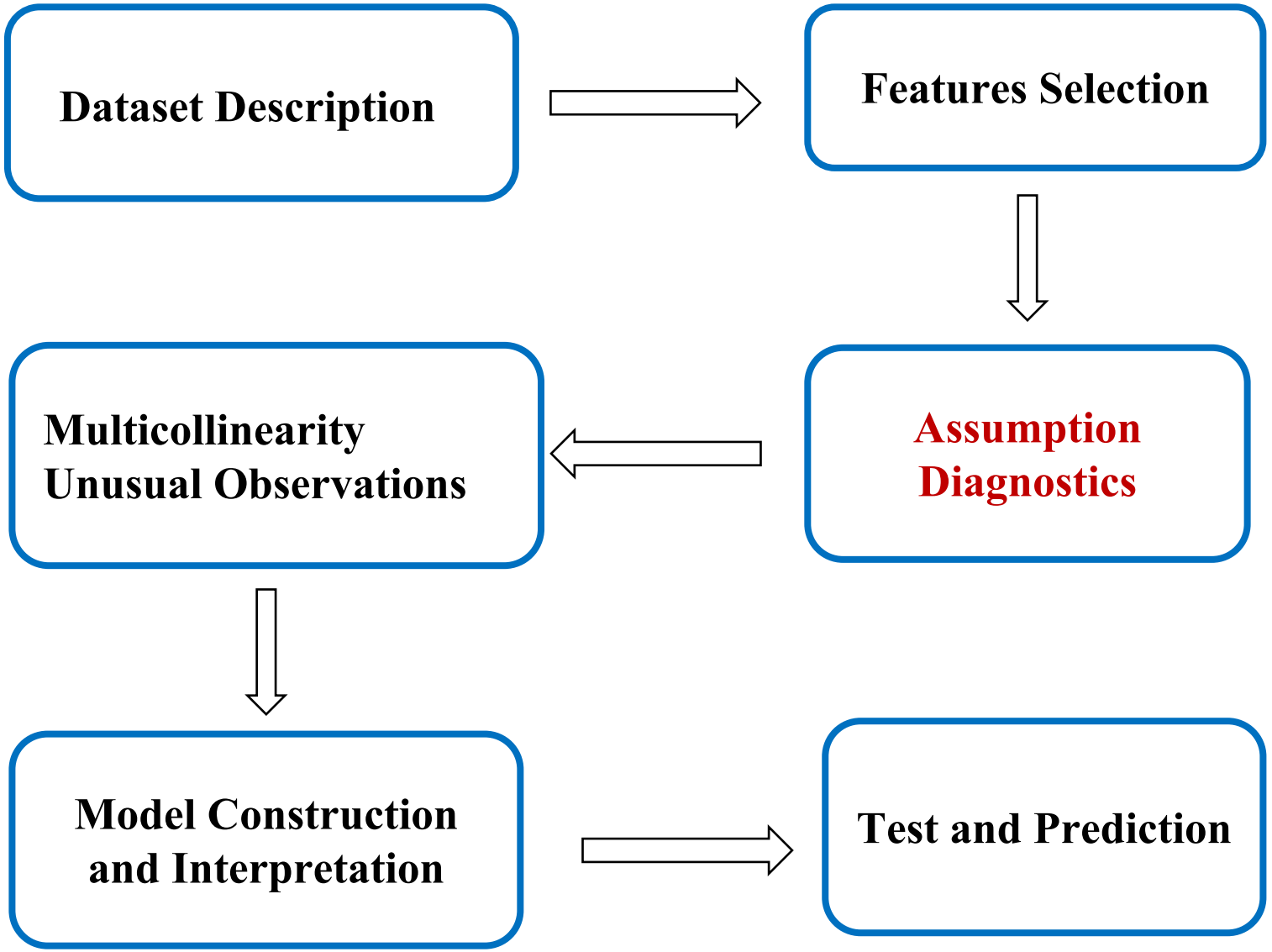
Feature Selection

- The chosen features are:

Rings ~ Shell_Weight + Shucked_Weight + Diameter + Whole_Weight + Gen_I + Viscera_Weight + Height + Gen_F

- Gender_M and length are eliminated.
- In order to better explain the realistic significance of the model, we decided to retain Gender_M.
- Thus, the features selected are:

Rings ~ Shell_Weight + Shucked_Weight + Diameter + Whole_Weight + Viscera_Weight + Height + Gen_I+ Gen_F+Gen_M


Dataset Description
Features Selection
Assumption Diagnostics
Multicollinearity Unusual Observations
Model Construction and Interpretation
Test and Prediction

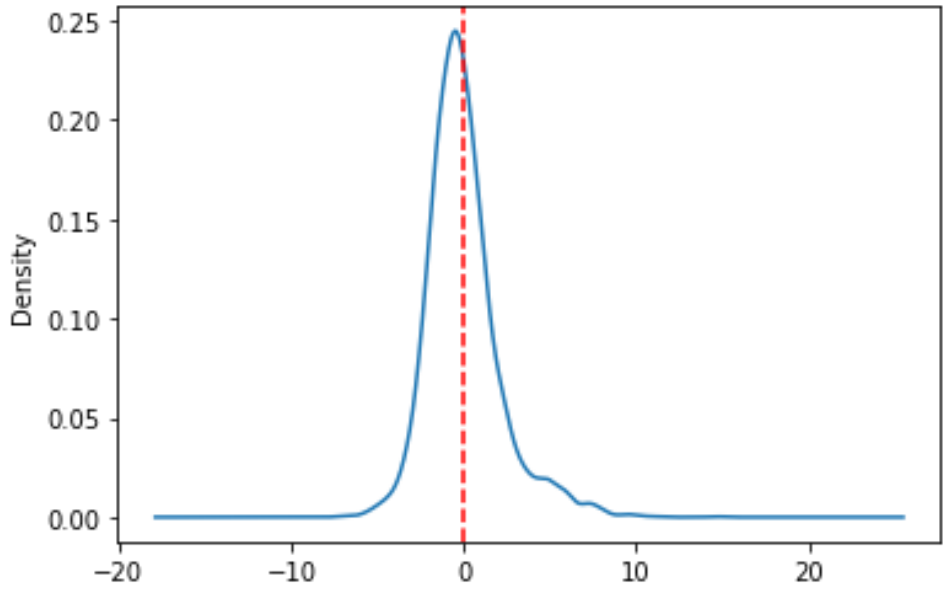
Assumption Diagnostics

- Normality Test
- Independence Test
- Homoscedasticity Test

Verifying Normality

Methods to Verify Normality

-Histogram of residuals

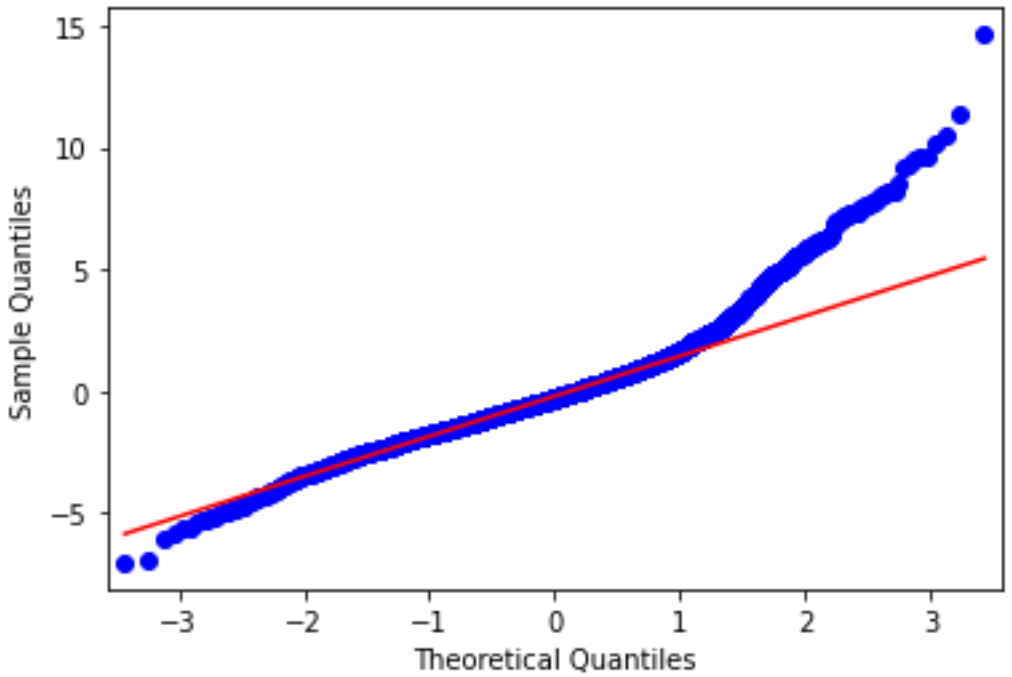


The plot shows that data is close to normal.

Verifying Normality

Methods to verify Normality

- Q-Q plot



Heavy-tailed

- Shapiro-Wilk TEST

- Null hypothesis:
  Residuals do follow normal distribution

- Alternative hypothesis:
  Residuals do NOT follows normal distribution

```
In [85]: import scipy.stats as stats
         stats.shapiro(residual)

Out[85]: (0.9253202676773071, 6.612140028832567e-38)

In [91]: stats.shapiro(residual)[1] < 0.05

Out[91]: True
```

Reject the null hypothesis at level 0.05, so the data is NOT normal.

Methods to Deal with Non-normality

- Box-Cox Transformation

Using maximum likelihood to estimate the skewness of the data, and converted to an approximate normal-terrestrial distribution using Eq.

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0; \\ (y_t^{\lambda} - 1)/\lambda & \text{otherwise.} \end{cases}$$

The result after implementing Box-Cox

```
In [92]: import scipy.stats as stats
         stats.shapiro(residual_box)[1] < 0.05
Out[92]: True
```

The Residuals are also non-normal after Box-Cox

```
In [95]: stats.shapiro(residual_box)[1] > stats.shapiro(residual)[1]
Out[95]: True
```

But it was improved

Methods to Deal with Non-normality

- Apply *log/sqrt* function to the response and refit the data

```
In [100]: import scipy.stats as stats
          stats.shapiro(lm_reg_log.resid)[1] < 0.05
Out[100]: True

In [102]: stats.shapiro(lm_reg_log.resid)[1] > stats.shapiro(residual)[1]
Out[102]: False
```

- log
  Become worse

```
In [104]: import scipy.stats as stats
          stats.shapiro(lm_reg_sqrt.resid)[1] < 0.05
Out[104]: True

In [105]: stats.shapiro(lm_reg_sqrt.resid)[1] > stats.shapiro(residual)[1]
Out[105]: False
```

- sqrt
  Become worse

Methods to Deal with Non-normality

After many methods applied to data, we found that it was hard to deal with non-normality in the data.

- Normal distribution test method is used to verify the most severe non-normality, which is not always suitable in practice.

- **Therefore, given the histogram, we have enough reason to believe that our data passes the normality test.**

Verifying Independence

Method to Test Residuals Independence

- Durbin-Watson Test

A value of DW = 2 indicates that there is no autocorrelation

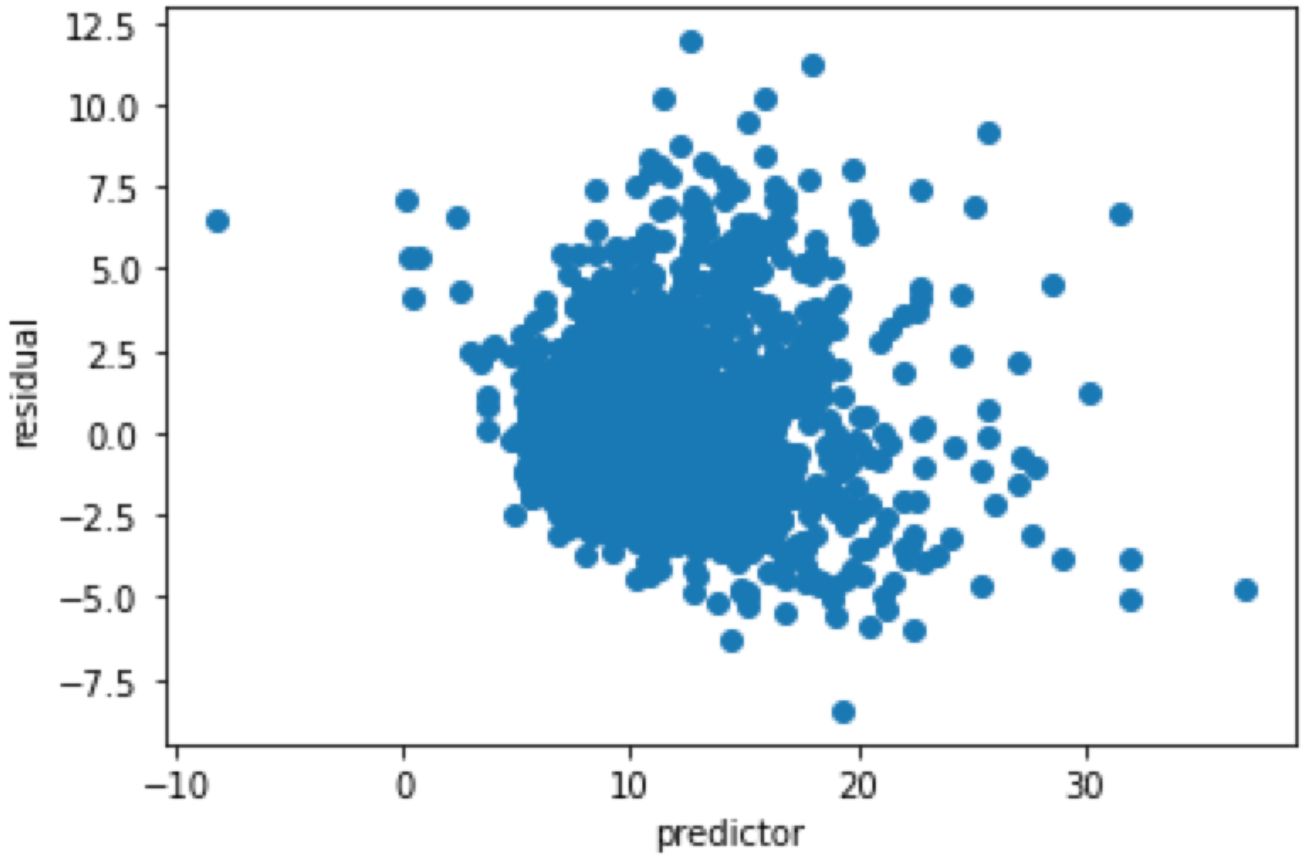
```
In [120]: from statsmodels.stats.stattools import durbin_watson as dwtest
          dwtest(resids=np.array(lm_reg.resid))

Out[120]: 1.9359455527363014
```

The residuals are independent

Verifying Constant Variance

Method to Verify Constant Variance

- Residual plot: fitted variable ~ residual



Without any pattern

Verifying Constant Variance

Method to Verify Constant Variance

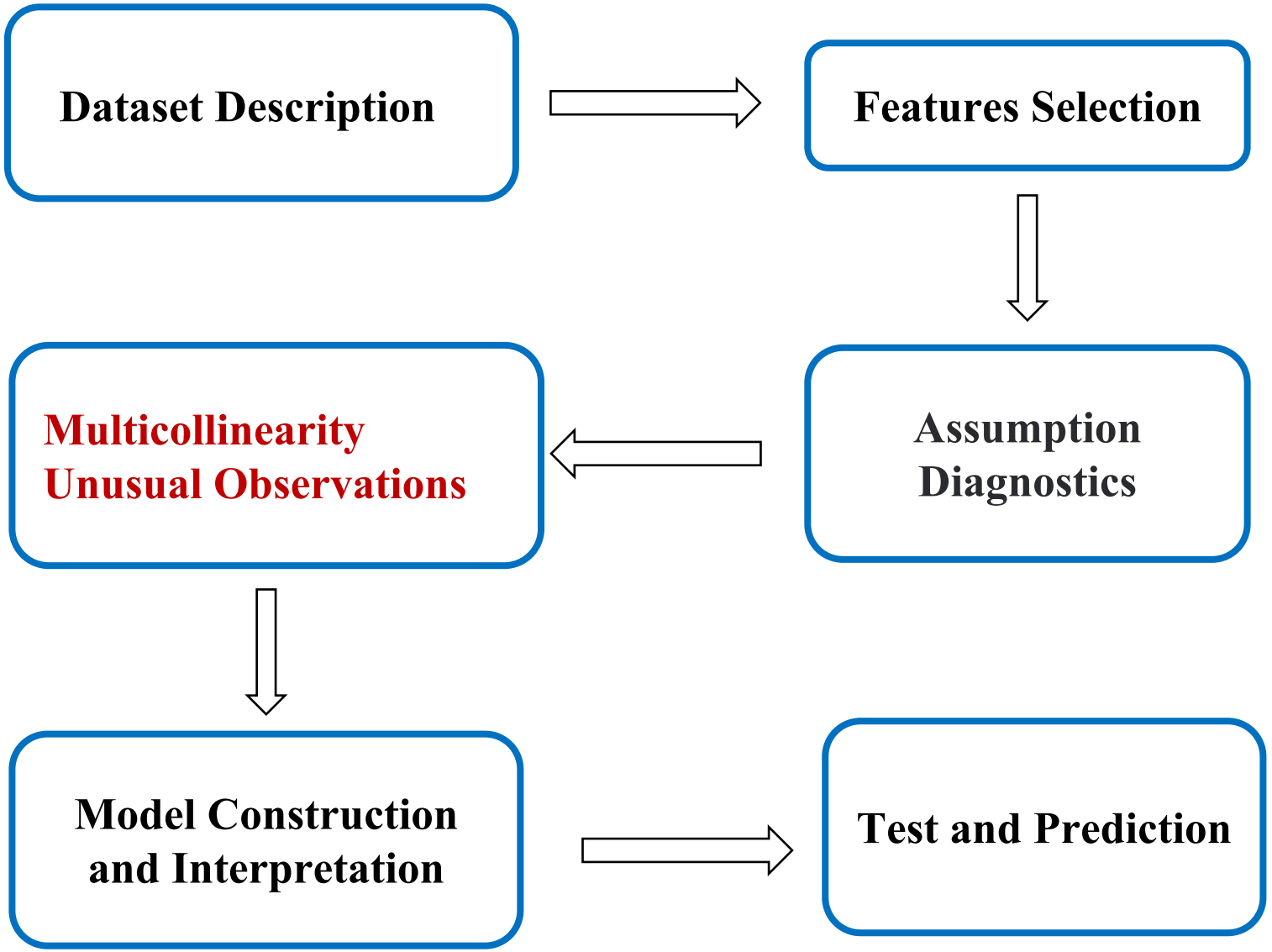
- Breusch-Pagan Testing

Check whether there is a linear relationship between the independent variable and the residual.
If so, it indicates that the variance may be inconsistent

```
In [246]: bp = sm.stats.diagnostic.het_breuschpagan(residual,lm_reg.model.exog)
          bp

Out[246]: (7.735577877362583, 0.6546474723549265, 0.8589247538465729, 0.561621640164987)
```

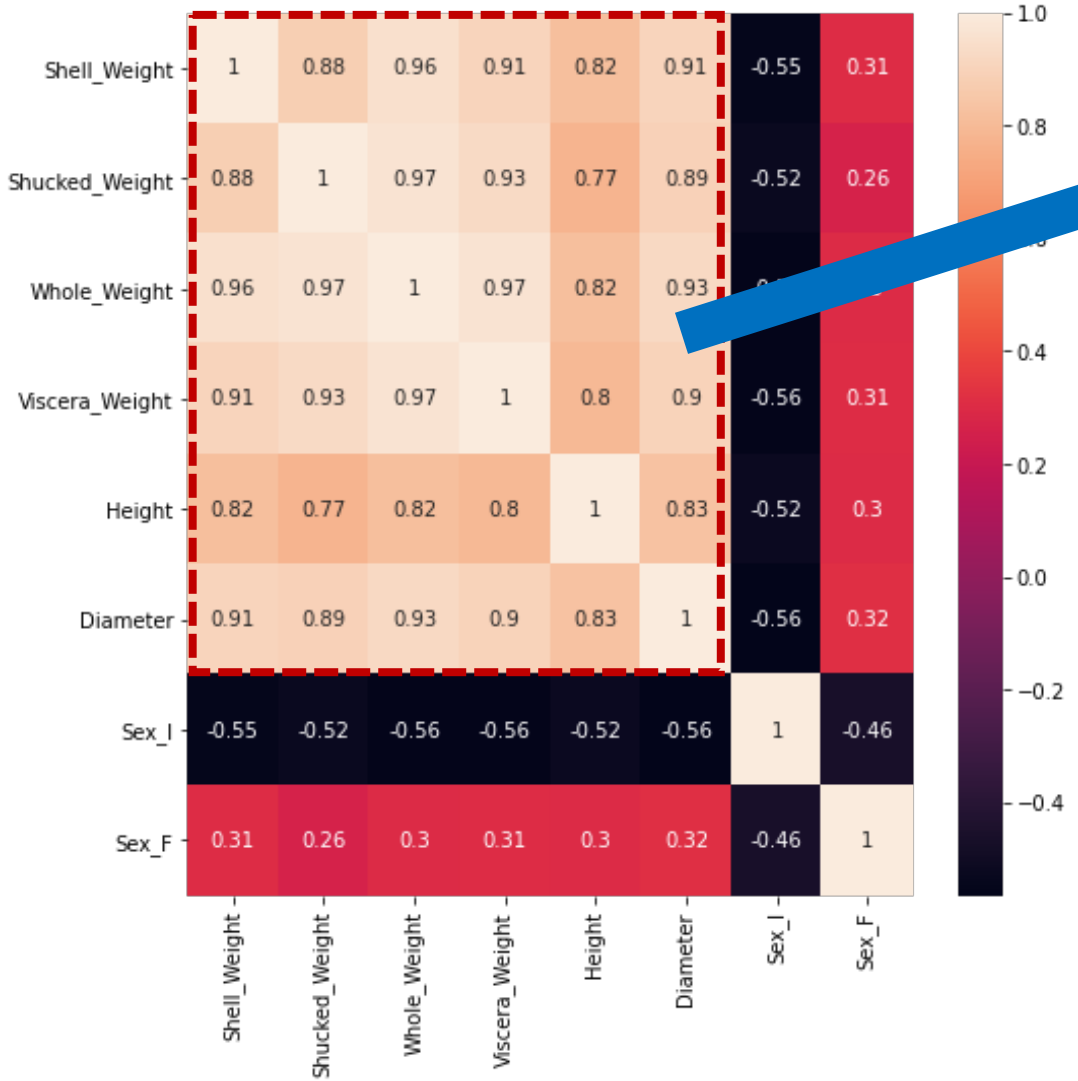
- The first one is: LM statistical value
- The second is: the P-value of the response is 0.65, which is higher than the significance level of 0.05. **Therefore, the null hypothesis(residuals' varience is constant) is accepted, that is, the residual variance is a constant**
- The third is the statistic value of F, which is used to test whether the square term of the residual is independent from the independent variable. If it is independent, it indicates the homogeneity of the variance of the residual
- The fourth is that the P-value corresponding to the F statistic is also higher than 0.05, so **the constancy of residuals' variance is verified.**

Residual TEST ITEM	method/index	value	conclusion
Normality test	histogram		Normal
	Q-Q plot	heavy-tailed	heavy-tailed distribution
	Shapiro-Wilk test	p-value < 0.05	*Abnormal*
			Can be treated as normal
Independence test	Durbin-Watson Test	Test statistics is close to 2	independent
Homoscedasticity Test	Plot: Residuals vs Fitted Values		constant
	BP testing	p-value of LM statistical value > 0.05	constant
		p-value of F statistic > 0.05	constant


Dataset Description
Features Selection
Assumption Diagnostics
Multicollinearity
Unusual Observations
Model Construction and Interpretation
Test and Prediction

Checking Predictor Multicollinearity

- First, check the heatmap between each independent variable



The corrs of these independent variables are very high, implying strong multicollinearity

Checking Predictor Multicollinearity

- Then, calculate the VIF of each independent variables.

Rule-of-thumb: Serious multicollinearity

- if average VIF of the $p - 1$ variables $\gg 1$;
- or if maximum VIF > 10.

```
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif
```

	VIF
0	62.333114
1	21.242202
2	28.386046
3	109.754856
4	17.289367
5	3.577715
6	8.355832
7	1.728211
8	1.295742

The VIFs much larger than 10

Methods to Deal with Multicollinearity

- Solution 1: Remove some features

We can easily find that

- the correlations between Whole_weight, Shucked_Weight , Viscera_Weight, Shell_Weight are large
- we can just retain 1 feature from the cluster (choose to retain **Whole_weight**)

The original model:

Rings ~ Whole_Weight + Shell_Weight + Shucked_Weight + Viscera_Weight + Diameter + Height + Gen_F + Gen_I

The new model:

Rings ~ Whole_Weight + Diameter + Height + Gen_F + Gen_I

```
# the original
lm_reg_original.rsquared
```

```
0.5424986379747911
```

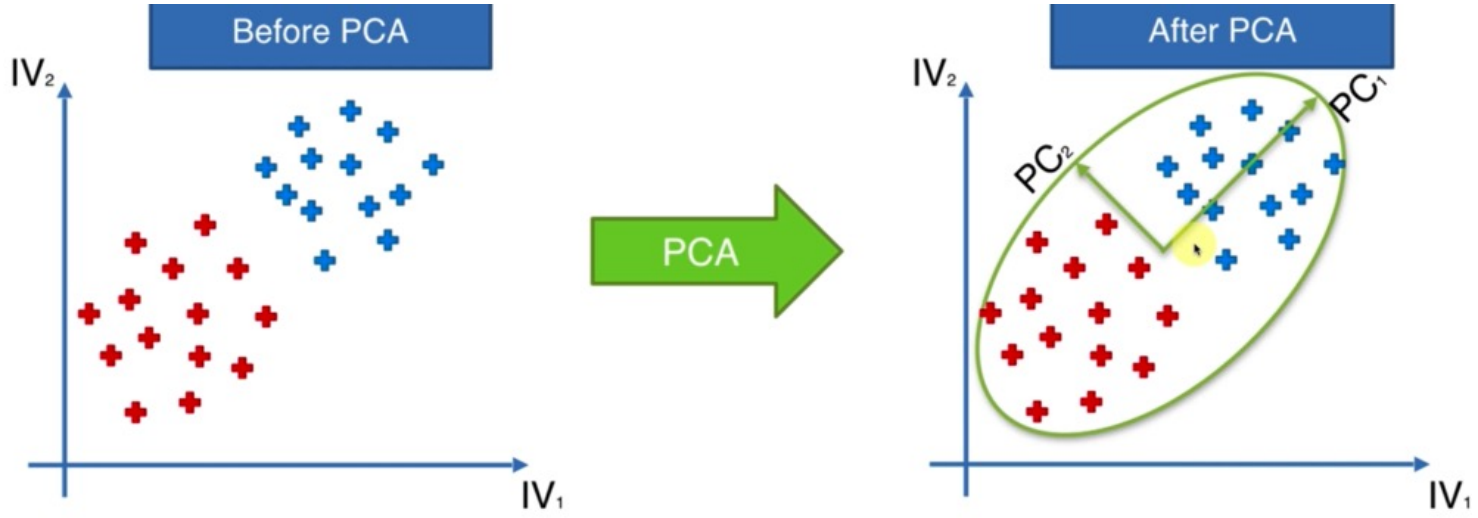
```
#the removed
lm_reg_remove.rsquared
```

```
0.3673094047535398
```

- After removing, the R-Square goes down rapidly, which means that the model's imitative effect becomes worse;
- Reason: A lot of information required for model fitting is lost after removing

Methods to deal with Multicollinearity

- **Solution 2: Principal Component Analysis (PCA) - Reduce the dimension of data**



2-dimension to 1-dimension

- **PCA is a black box, because it involves matrix transformations too many times', it will lose the interpretability of the model**
- **So when we go after the interpretability of the model, PCA will NOT be a ideal method to deal with non-normality**

Methods to deal with Multicollinearity

- **Solution 2: Principal Component Analysis (PCA) - Reduce the dimension of data**

PCA dimension's decision method: MLE
Minka's MLE is used to guess the dimension.

The original model:

Rings ~ Whole_Weight + Shell_Weight + Shucked_Weight + Viscera_Weight + Diameter + Height + Gen_F + Gen_I

8-dimension to 6-dimension

Model after PCA:

Rings ~ pca_weight_1 + pca_weight_2 + pca_weight_3 + pca_weight_4 + pca_weight_5 + pca_weight_6

R-Square does not change too much

```
# the original
lm_reg_original.rsquared
```

```
0.5424986379747911
```

```
# model after PCA
lm_reg_pca.rsquared
```

```
0.5283808381275735
```

VIF changes a lot

	VIF
0	62.333114
1	21.242202
2	28.386046
3	109.754856
4	17.289367
5	3.577715
6	8.355832
7	1.728211
8	1.295742

	VIF
0	1.0
1	1.0
2	1.0
3	1.0
4	1.0
5	1.0
6	1.0

Methods to deal with Multicollinearity

- Solution 3: Ridge/Lasso Regression

- Linear regression fitted the data with a linear function, calculated the cost with mean square error (MSE), and then found a set of weights that minimized MSE using gradient descent;
- Lasso and Ridge regression are to add L1 and L2 regularization respectively on the basis of standard linear regression

$$Lasso : J = \frac{1}{n}\Sigma_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda||\beta||_1$$

$$Ridge : J = \frac{1}{n}\Sigma_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda||\beta||_2^2$$

We choose Ridge in this Section

Checking Unusual Observations

- Leverage Point

- Outlier

- Influential observations

Check Leverage Point

- hatvalue

- The average leverage for each data point is p/n.
- Rule of thumb: Leverages larger than $2p/n$ are considered "large", and the corresponding observations should be looked at more closely.

```
hat_value_point = 2*sum(h)/x_train.shape[0]
hat_value_point
```

```
0.004788985333732418
```

```
OLSInfluence(model).summary_frame()[OLSInfluence(model).summary_frame()['hat_diag']>hat_value_point][['hat_diag']]
```

	hat_diag
159	0.006435
593	0.005283
753	0.005444
375	0.005186
1757	0.007776
...	...
2199	0.007815
660	0.005462
2177	0.006476
1204	0.007686
3149	0.013159

211 rows × 1 columns

If hatvalue>0.0047, take it as leverage point

211/3814 points were set as leverage point

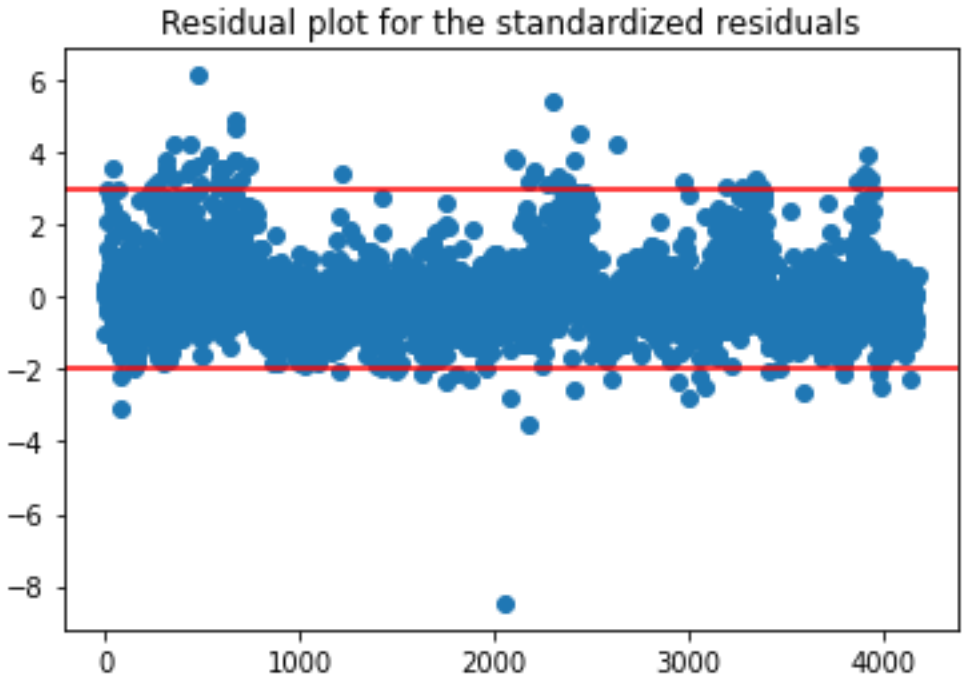
Check Leverage Point

- Standardized Residuals

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}.$$

If model assumptions are correct, $\mathrm{Var}(r_i) = 1$, and $\mathrm{Cor}(r_i, r_j)$ tend to be small. Thus, standardized residuals are sometimes preferred in residual plots.

residual plot – standard residual



44/3814 points were set as leverage point by comparing standard residual

SUMMARY: combine 2 methods
211/3814 points was set as leverage point

Check Outlier

- Jackknife residuals

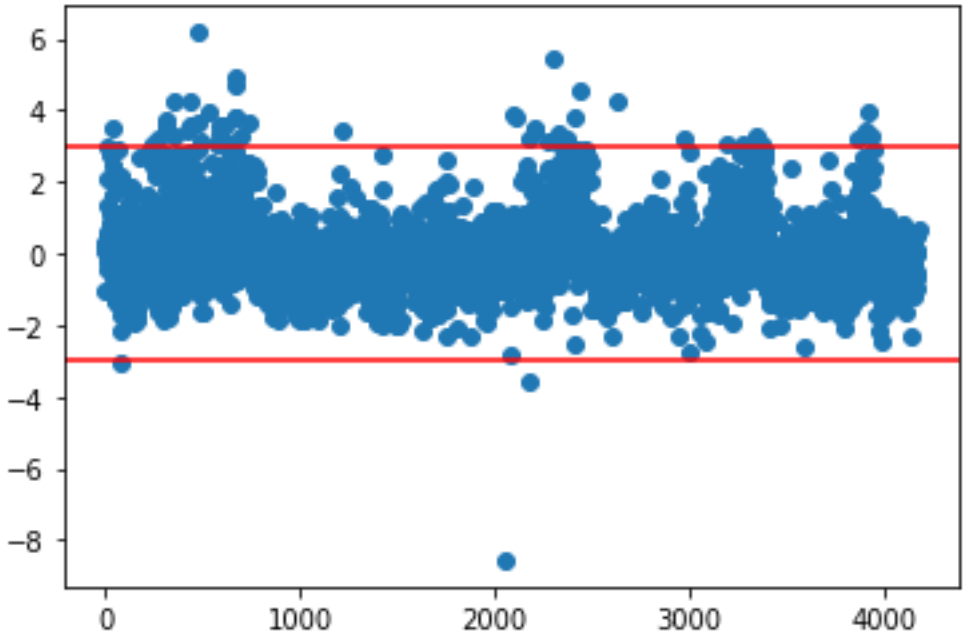
The jackknife residual for the i-th case is defined as

$$t_i = \frac{y_i - \hat{y}_{(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(1 + \mathbf{x}_i^\top (\mathbf{X}_{(i)}^\top \mathbf{X}_{(i)})^{-1}\mathbf{x}_i)}} = \frac{\hat{\epsilon}_i}{\sqrt{\hat{\sigma}^2_{(i)}(1 - h_{ii})}} = r_i \left(\frac{n-p-1}{n-p-r_i^2}\right)^{1/2}$$

where $\mathbf{X}_{(i)}$ is the design matrix excluding the i-th case.

- Generally, when the sample size is **several hundred**, the points with studization residuals **greater than 2** are regarded as strong influence points
- When the sample size is **thousands**, the points with studization residuals **greater than 3** are relatively large influence points.

Jackknife residual plot



54/3814 points were set as outliers by comparing Jackknife Residual

Check Influential Observations

- Cook's Distance

Cook's *D*-statistic or Cook's distance:

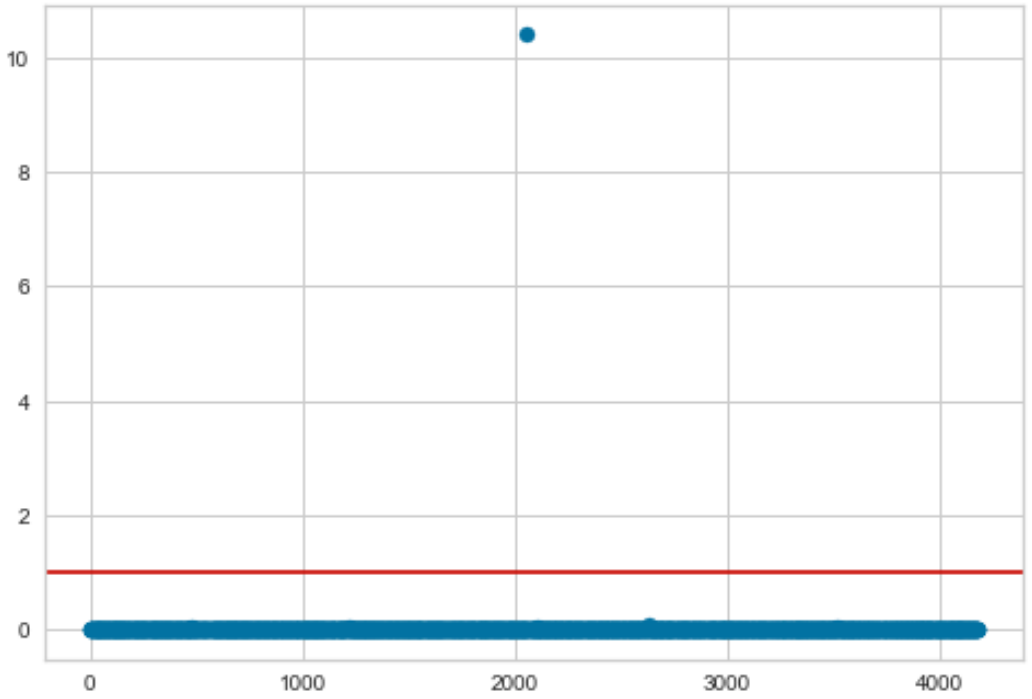
$$D_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})^\top (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{p\hat{\sigma}^2} = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}.$$

where $\mathbf{y}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$ is the jackknife fit, r_i the standardized residual, and h_i the leverage.

Absolutely large residual and large leverage will result in large influence.

Usually $D_i > 1$ suggests an influential point.

Cook's Dist plot



1/3814 points was set as influential point by comparing Cook's Distance

Unusual Observations

Unusual Observations	method	point number
Leverage Point	hatvalue	211
	Standardized Residuals	44
Outlier	**Jackknife residuals**	54
Influential observations	**Cook's Distance**	1

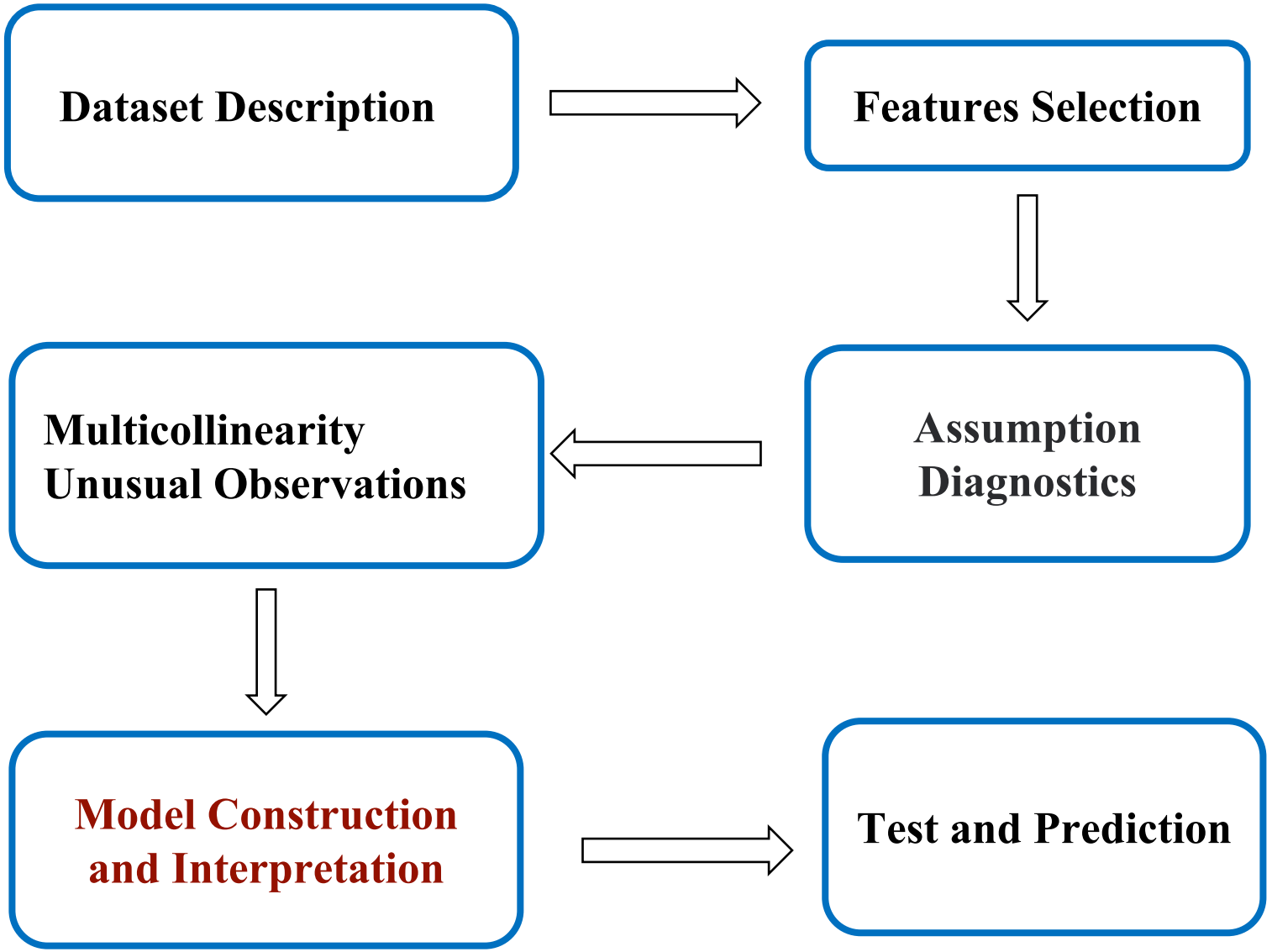
Unusual Observations

```
model_summary[(model_summary['hat_diag']>hat_value_point) | (abs(model_summary['student_resid'])>3)
              | (model_summary['cooks_d']>1)][['hat_diag','student_resid','cooks_d']]
```

	hat_diag	student_resid	cooks_d
3924	0.001720	3.940710	0.003330
159	0.006435	-1.547448	0.001938
593	0.005283	1.775728	0.002092
753	0.005444	0.194842	0.000026
375	0.005186	2.416031	0.003798
...	...	...	...
660	0.005462	1.584071	0.001722
2177	0.006476	-0.941334	0.000722
572	0.003250	3.054409	0.003793
1204	0.007686	0.231176	0.000052
3149	0.013159	2.482214	0.010254

252 rows × 3 columns

- **252/3814 points were set as unusual observations;**
- **Solution: delete them**

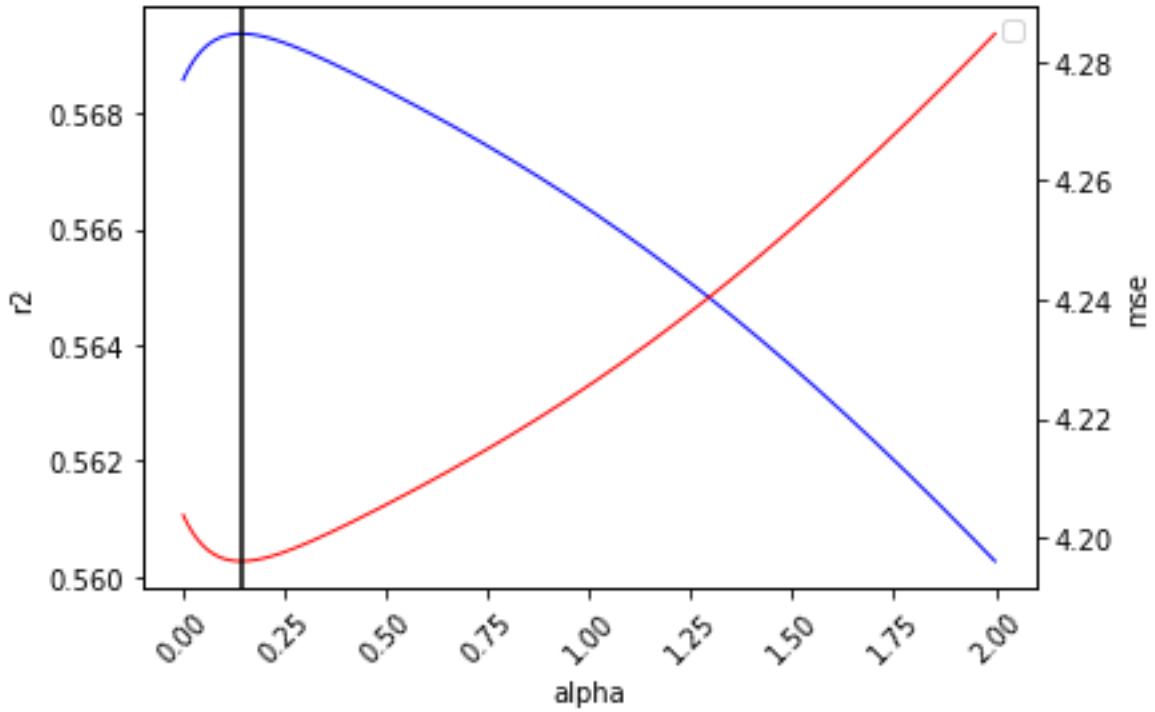

Dataset Description
Features Selection
Assumption Diagnostics
Multicollinearity
Unusual Observations
Model Construction and Interpretation
Test and Prediction

Tuning parameter

$$Ridge : J = \frac{1}{n}\Sigma_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda||\beta||_2^2$$

We need to find the point that maximizes R-squared value and minimizes MSE.

we choose the point which value of λ is 0.144



Interpretation

intercept	Shell_Weight	Shucked_Weight	Whole_Weight	Viscera_Weight	Height	Diameter	Gen_F	Gen_I	Gen_M
3.81	**11.33**	**-19.36**	**7.54**	**-8.62**	**19.45**	**6.93**	**0.28**	**-0.61**	**0.33**

The Ridge regression Model:

$$\begin{aligned} Rings = {} & 11.33 Shell\_weight - 19.36 Shucked\_weight + 7.54 Whole\_weight - 8.62 Viscera\_weight \\ & + 19.45 Height + 6.93 Diametre \\ & + 0.28 Gender\_F - 0.61 Gender\_I + 0.33 Gender\_M \\ & + 3.81 \end{aligned}$$

Interpretation

intercept	Shell_Weight	Shucked_Weight	Whole_Weight	Viscera_Weight	Height	Diameter	Gen_F	Gen_I	Gen_M
3.81	**11.33**	**-19.36**	**7.54**	**-8.62**	**19.45**	**6.93**	**0.28**	**-0.61**	**0.33**

Weight

Qualitatively speaking, the number of abalone rings is positively correlated with the height and diameter, which is in line with intuition.

By looking at the four variables related to weight, we can conclude that the older the abalone, the greater the total weight, the greater the proportion of the shell weight, and conversely, the smaller the proportion of the viscera weight.

Quantitatively speaking, when all other things being equal, adult male abalone has 0.05 more rings than female. There is no point in directly comparing the number of rings between adult and infant abalone, because the dummy variable itself represents a distinct age difference.

Interpretation

intercept	Shell_Weight	Shucked_Weight	Whole_Weight	Viscera_Weight	Height	Diameter	Gen_F	Gen_I	Gen_M
3.81	**11.33**	**-19.36**	**7.54**	**-8.62**	**19.45**	**6.93**	**0.28**	**-0.61**	**0.33**

Length

Qualitatively speaking, the number of abalone rings is positively correlated with the height and diameter, which is in line with intuition.

By looking at the four variables related to weight, we can conclude that the older the abalone, the greater the total weight, the greater the proportion of the shell weight, and conversely, the smaller the proportion of the viscera weight.

Quantitatively speaking, when all other things being equal, adult male abalone has 0.05 more rings than female. There is no point in directly comparing the number of rings between adult and infant abalone, because the dummy variable itself represents a distinct age difference.

Interpretation

intercept	Shell_Weight	Shucked_Weight	Whole_Weight	Viscera_Weight	Height	Diameter	Gen_F	Gen_I	Gen_M
3.81	**11.33**	**-19.36**	**7.54**	**-8.62**	**19.45**	**6.93**	**0.28**	**-0.61**	**0.33**

Gender

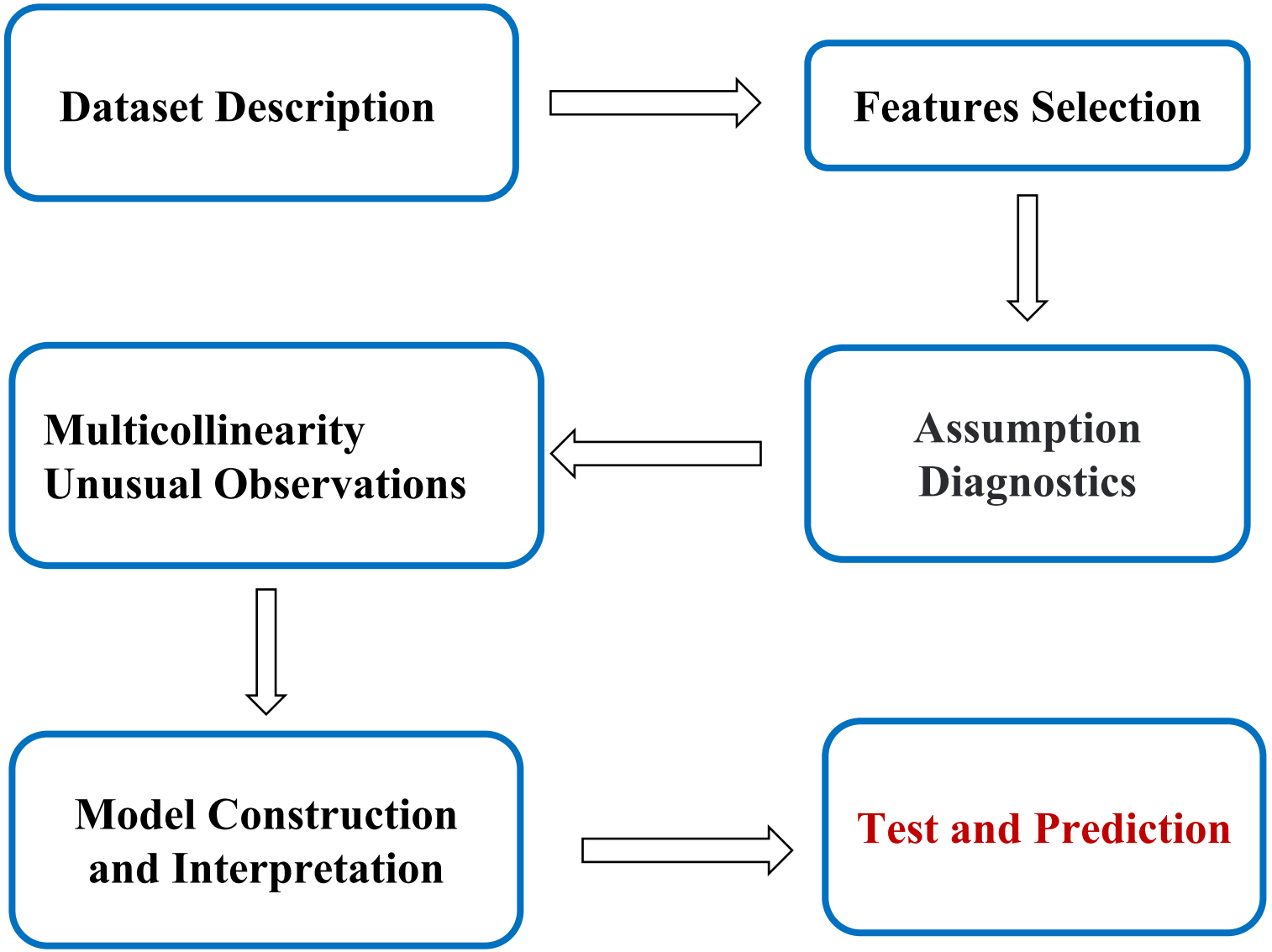
Qualitatively speaking, the number of abalone rings is positively correlated with the height and diameter, which is in line with intuition.

By looking at the four variables related to weight, we can conclude that the older the abalone, the greater the total weight, the greater the proportion of the shell weight, and conversely, the smaller the proportion of the viscera weight.

Quantitatively speaking, when all other things being equal, adult male abalone has 0.05 more rings than female. There is no point in directly comparing the number of rings between adult and infant abalone, because the dummy variable itself represents a distinct age difference.

Interpretation

```
Residuals:
    Min       1Q  Median        3Q       Max
-3.2086  8.2201  9.6301  11.2589  32.0346


Coefficients:
                  Estimate    Std. Error  t value   p value
_intercept        3.584582  2.329213e+06   0.0000  0.999999
Shell_Weight      9.746884  1.309032e+00   7.4459  0.000000
Shucked_Weight  -20.641312  8.648790e-01 -23.8661  0.000000
Whole_Weight      8.787933  5.190520e-01  16.9307  0.000000
Viscera_Weight  -10.665626  1.593416e+00  -6.6936  0.000000
Height           24.365884  1.693545e+00  14.3875  0.000000
Diameter          6.250309  1.188843e+00   5.2575  0.000000
Sex_F             0.266598  2.329213e+06   0.0000  1.000000
Sex_I            -0.594187  2.329213e+06  -0.0000  1.000000
Sex_M             0.327589  2.329213e+06   0.0000  1.000000
---
R-squared:  0.51717,    Adjusted R-squared:  0.51554
F-statistic: 316.93 on 9 features
```


Dataset Description
Features Selection
Assumption Diagnostics
Multicollinearity Unusual Observations
Model Construction and Interpretation
Test and Prediction

Test and prediction

Difference < 1	Difference < 2
46.11%	74.37%

Thank you!