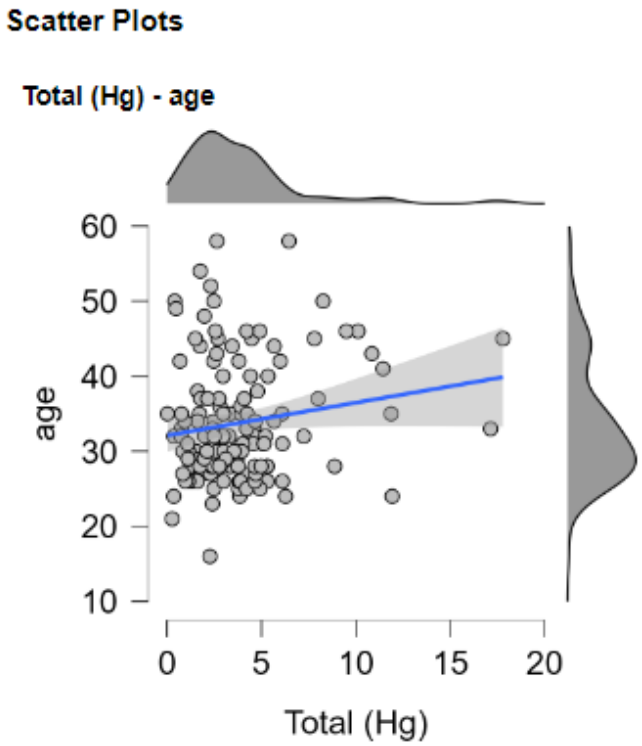
Emmanuel Pedernal

1. **Determine the relationship between each of the numerical independent variables (except the categorical) and the dependent variable by generating scatterplots for each.**

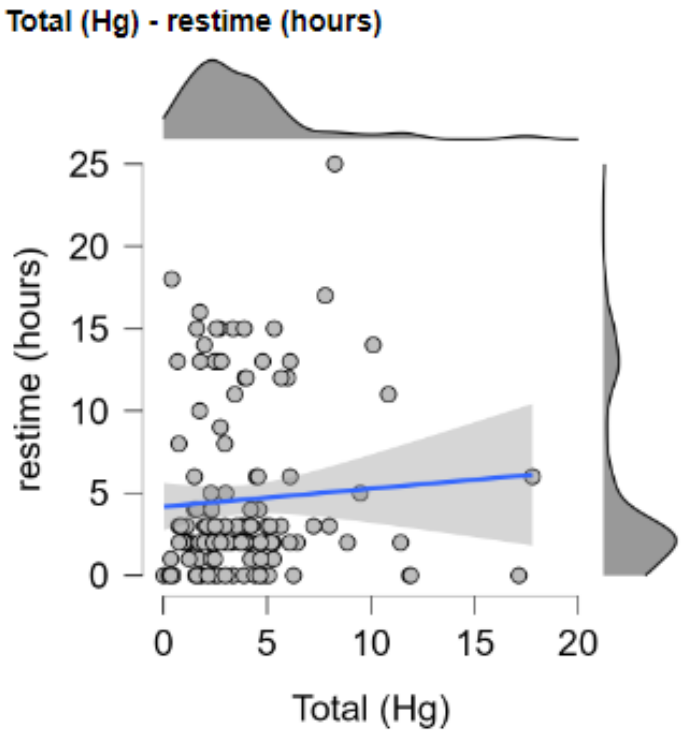
Pearson's Correlations

Variable		Total (Hg)	age	weight (kg)	number of fish meals per week	height (cm)	restime (hours)
1. Total (Hg)	Pearson's r	—					
	p-value	—					
2. age	Pearson's r	0.160	—				
	p-value	0.064	—				
3. weight (kg)	Pearson's r	0.409	0.046	—			
	p-value	< .001	0.598	—			
4. number of fish meals per week	Pearson's r	0.303	0.257	0.040	—		
	p-value	< .001	0.003	0.647	—		
5. height (cm)	Pearson's r	0.192	−0.002	0.302	−0.044	—	
	p-value	0.026	0.980	< .001	0.609	—	
6. restime (hours)	Pearson's r	0.062	0.583	0.105	0.189	−0.054	—
	p-value	0.474	< .001	0.225	0.028	0.535	—



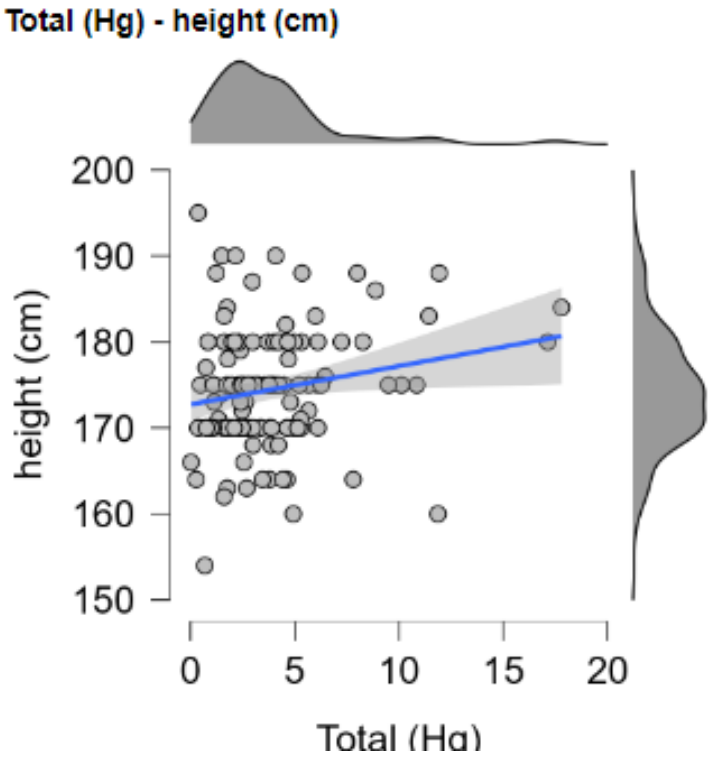
age has positive but weak relationship with dependent variable (Total Hg)

Pearson's r = 0.160 and p-value = 0.064



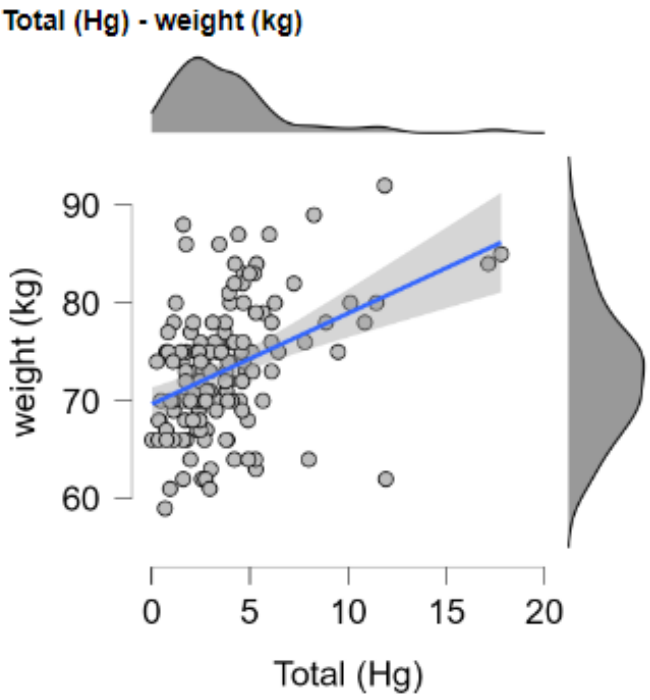
restime (hours) has positive but weak relationship with dependent variable (Total Hg)

Pearson's r = 0.062 and p-value = 0.474



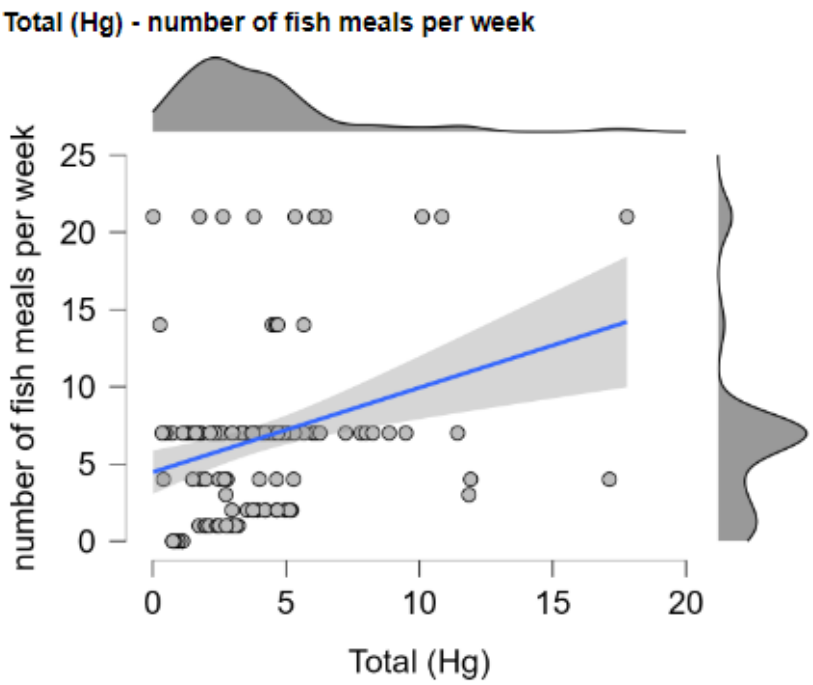
Height has positive but weak relationship with dependent variable (Total Hg)

Pearson's r = 0.192 and p-value = 0.026



Weight (kg) has positive and has moderate relationship with dependent variable (Total Hg)

Pearson's r = 0.409 and p-value = 0.001



Number of fish meals per week has positive and has moderate relationship with dependent variable (Total Hg)

Pearson's r = 0.303 and p-value = 0.001

The results show a statistically significant positive correlation with the dependent variable (Total Mercury) with weight and number of fish meals per week. The data indicates that a person who consumes more fish and higher body weight has higher chance of having high levels of mercury than those who eat less and weigh less.

2. **Generate the multiple regression models using full model, forward stepwise, and backward stepwise (include the variable owned_boat.**

Full Model LR

Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC
M_0	0.000	0.000	0.000	2.939	677.146	682.957
M_1	0.521	0.271	0.237	2.566	646.404	669.646

Note. M_1 includes age, restime (hours), height (cm), weight (kg), number of fish meals per week, owned_boat

Forward LR

Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC
M_0	0.000	0.000	0.000	2.939	677.146	682.957
M_1	0.409	0.167	0.161	2.691	654.398	663.114
M_2	0.500	0.250	0.238	2.564	642.349	653.970

Coefficients Covariance Matrix

Model		weight (kg)	number of fish meals per week
M_1	weight (kg)	0.001	
M_2	weight (kg)	0.001	-5.518×10^{-5}
	number of fish meals per week		0.002

Note. The intercept model is omitted, as no meaningful information can be shown.

Backward LR ▼

Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC
M_0	0.521	0.271	0.237	2.566	646.404	669.646
M_1	0.519	0.269	0.241	2.560	644.774	665.111
M_2	0.513	0.263	0.240	2.561	643.976	661.408
M_3	0.505	0.255	0.238	2.566	643.452	657.979
M_4	0.500	0.250	0.238	2.564	642.349	653.970

Coefficients Covariance Matrix

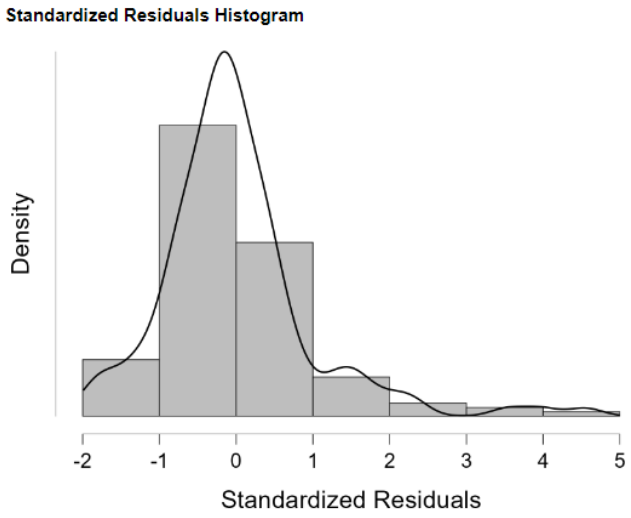
Model		age	restime (hours)	height (cm)	weight (kg)	number of fish meals per week	owned_boat
M_0	age	0.001	−0.001	-6.287×10^{-5}	4.318×10^{-5}	-3.020×10^{-4}	0.002
	restime (hours)		0.003	1.769×10^{-4}	-2.215×10^{-4}	-7.135×10^{-5}	−0.003
	height (cm)			0.001	-3.770×10^{-4}	7.777×10^{-5}	6.603×10^{-6}
	weight (kg)				0.001	-5.699×10^{-5}	-3.288×10^{-4}
	number of fish meals per week					0.002	−0.002
	owned_boat						0.207
M_1	age	0.001	−0.001	-6.263×10^{-5}	4.682×10^{-5}	-2.779×10^{-4}	
	restime (hours)		0.003	1.761×10^{-4}	-2.250×10^{-4}	-9.784×10^{-5}	
	height (cm)			0.001	-3.751×10^{-4}	7.744×10^{-5}	
	weight (kg)				0.001	-5.975×10^{-5}	
	number of fish meals per week					0.002	
M_2	age	0.001	−0.001		2.679×10^{-5}	-2.741×10^{-4}	
	restime (hours)		0.003		-1.688×10^{-4}	-1.096×10^{-4}	
	weight (kg)				0.001	-3.499×10^{-5}	
	number of fish meals per week				-3.499×10^{-5}	0.002	
M_3	age	8.170×10^{-4}			-3.508×10^{-5}	-3.152×10^{-4}	
	weight (kg)				0.001	-4.170×10^{-5}	
	number of fish meals per week				-4.170×10^{-5}	0.002	
M_4	weight (kg)				0.001	-5.518×10^{-5}	
	number of fish meals per week				-5.518×10^{-5}	0.002	

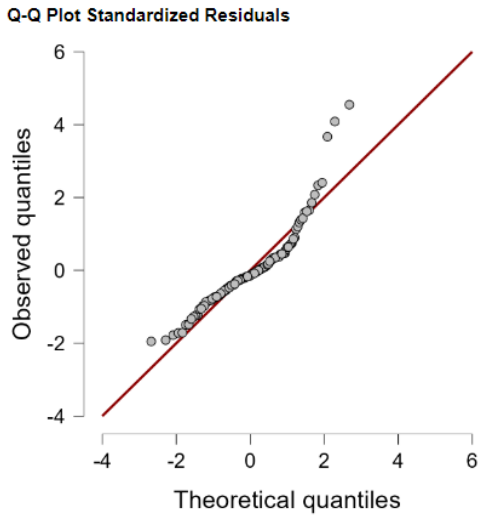
3. **Compare the 3 models, select and interpret the best model.**

The best models are LR **Forward and Backward** with the same Adjusted R^2 of 0.238 and RMSE of 2.564. Both models use fewer (2) predictors instead of all variables resulted in more interpretable and efficient models in addition having only few variables means it's easier to understand/communicate and reduce risk of multicollinearity.

4. **Do the model diagnostics/assumptions checking of the residuals by examining/testing the normality, homoscedasticity, independence, and linearity/multicollinearity.**

Normality Test (FULL MODEL)





Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC	Durbin-Watson		
							Autocorrelation	Statistic	p
M_0	0.000	0.000	0.000	2.939	677.146	682.957	0.302	1.387	< .001
M_1	0.521	0.271	0.237	2.566	646.404	669.646	0.191	1.616	0.016

Note. M_1 includes age, restime (hours), height (cm), weight (kg), number of fish meals per week, owned_boat

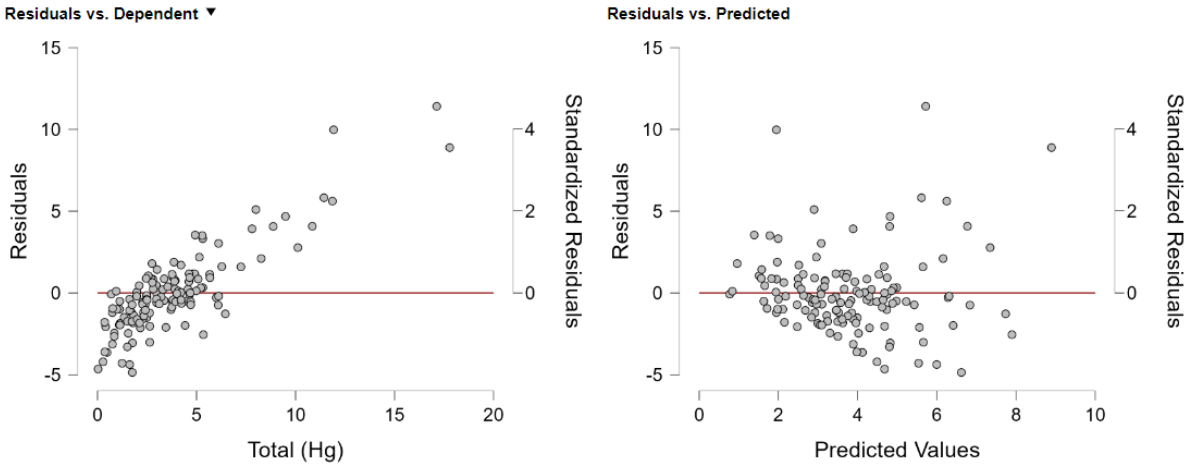
Descriptive Statistics

Descriptive Statistics

	Total (Hg)
Std. Deviation	2.939
Shapiro-Wilk	0.816
P-value of Shapiro-Wilk	< .001

The model failed in Normality test the Shapiro-wilk P value is less than .001 which means we'll reject the null hypothesis that residuals follow normal distribution.

Homoscedasticity



Upon visual assessment, there is a change in the spread or variance of the data points, hence the model does not satisfy homoscedasticity

Linearity

The plot Residuals vs. Predicted does not show linearity thus it failed the test. This could mean that we might need to log transform our model to fully capture the relationship between dependent and independent variables.

Independence

Full Model LR

Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC	Durbin-Watson Autocorrelation	Durbin-Watson Statistic	Durbin-Watson p
M_0	0.000	0.000	0.000	2.939	677.146	682.957	0.302	1.387	< .001
M_1	0.521	0.271	0.237	2.566	646.404	669.646	0.191	1.616	0.016

Note. M_1 includes age, restime (hours), height (cm), weight (kg), number of fish meals per week, owned_boat

Based from Durbin-Watson test, the DW statistic of 1.616, suggests positive auto correlation in residuals which means they are not independent.

Multicollinearity

Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Intercept)	age	restime (hours)	height (cm)	weight (kg)	number of fish meals per week	owned_boat
M_1	1	5.632	1.000	0.000	0.001	0.007	0.000	0.000	0.008	0.009
	2	0.554	3.188	0.000	0.001	0.147	0.000	0.000	0.009	0.746
	3	0.454	3.524	0.000	0.001	0.494	0.000	0.001	0.007	0.206
	4	0.328	4.146	0.000	0.002	0.013	0.000	0.001	0.947	0.002
	5	0.027	14.466	0.002	0.951	0.301	0.003	0.028	0.026	0.036
	6	0.005	34.248	0.062	0.043	0.037	0.045	0.956	0.001	0.001
	7	7.416×10^{-4}	87.147	0.935	0.002	0.001	0.952	0.014	0.002	0.001

Note. The intercept model is omitted, as no meaningful information can be shown.

Collinearity Statistics	
Tolerance	VIF
0.621	1.610
0.638	1.569
0.897	1.115
0.891	1.122
0.920	1.087
0.968	1.033

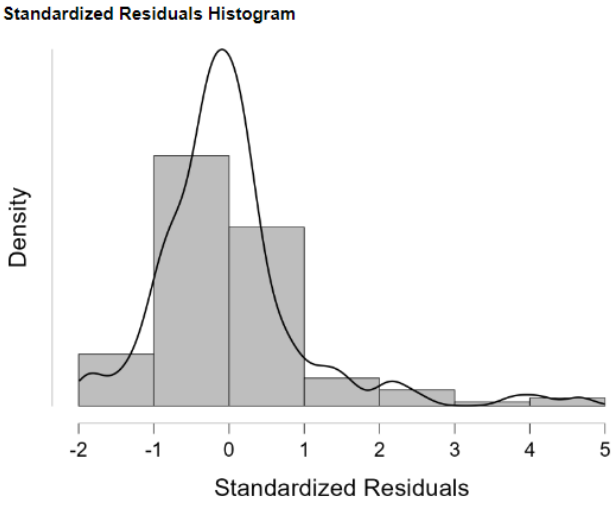
lke, Bayarri, & Berger,

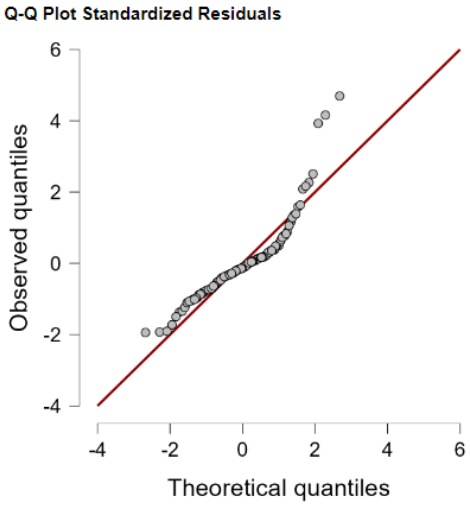
According to the condition index, there are variables greater than 30 hence the model failed the multicollinearity test.

Summary:

Model Diagnostics/Assumptions	**Results**
Normality (Histogram/Formal Test)	**Failed**
Homoscedasticity (Variance Plot)	**Failed**
Linearity (Plots)	**Failed**
Independence (DW Test)	**Failed**
Multicollinearity (Collinearity Diagnostics	**Failed**

Normality Test (Forward Pass)





Forward LR ▾

Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC	Durbin-Watson		
							Autocorrelation	Statistic	p
M_0	0.000	0.000	0.000	2.939	677.146	682.957	0.302	1.387	< .001
M_1	0.409	0.167	0.161	2.691	654.398	663.114	0.355	1.286	< .001
M_2	0.500	0.250	0.238	2.564	642.349	653.970	0.263	1.473	0.002

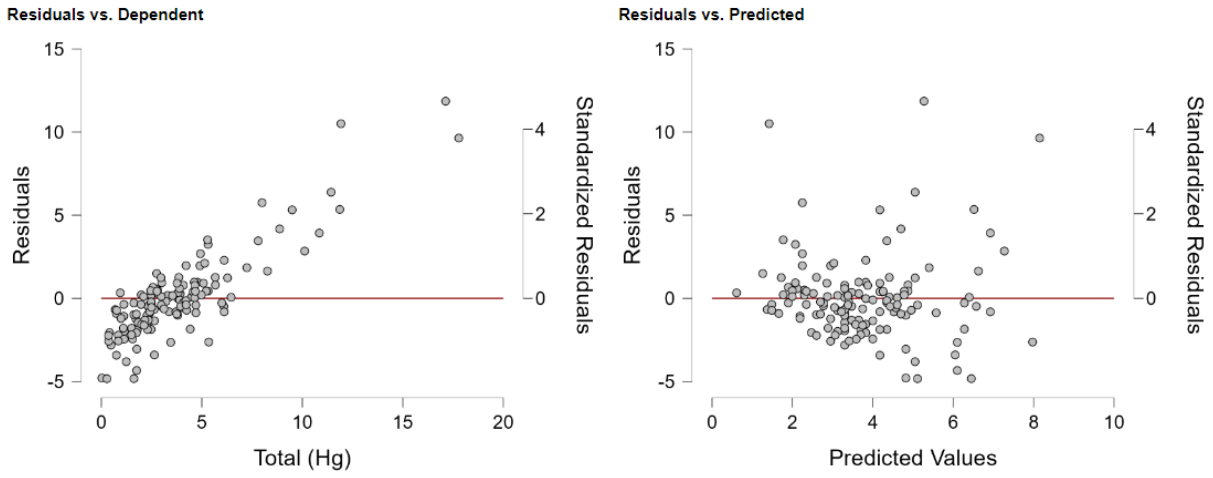
Descriptive Statistics

Descriptive Statistics

	Total (Hg)
Std. Deviation	2.939
Shapiro-Wilk	0.816
P-value of Shapiro-Wilk	< .001

The model failed in Normality test the Shapiro-wilk P value is less than .001 which means we'll reject the null hypothesis that residuals follow normal distribution.

Homoscedasticity



Upon visual assessment, there is a change (shape) in the spread or variance of the data points, hence the model does not satisfy homoscedasticity

Linearity

The plot Residuals vs. Predicted does not show linearity thus it failed the test. This could mean that we might need to log transform our model to fully capture the relationship between dependent and independent variables.

Independence

Forward LR ▾

Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC	Durbin-Watson		
							Autocorrelation	Statistic	p
M_0	0.000	0.000	0.000	2.939	677.146	682.957	0.302	1.387	< .001
M_1	0.409	0.167	0.161	2.691	654.398	663.114	0.355	1.286	< .001
M_2	0.500	0.250	0.238	2.564	642.349	653.970	0.263	1.473	0.002

Based from Durbin-Watson test, the DW statistic of 1.473, suggests positive auto correlation in residuals which means they are not independent.

Multicollinearity

Collinearity Diagnostics

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Intercept)	weight (kg)	number of fish meals per week
M_1	1	1.996	1.000	0.002	0.002	
	2	0.004	22.051	0.998	0.998	
M_2	1	2.703	1.000	0.001	0.001	0.043
	2	0.292	3.040	0.004	0.004	0.957
	3	0.004	25.667	0.995	0.995	0.000

Note. The intercept model is omitted, as no meaningful information can be shown.

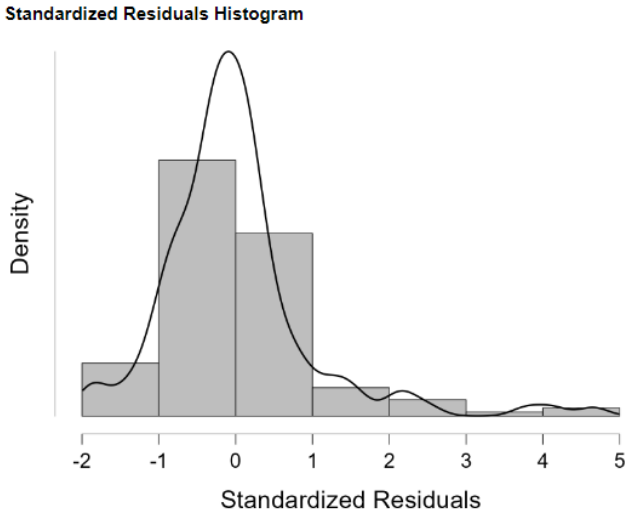
Collinearity Statistics	
Tolerance	VIF
1.000	1.000
0.998	1.002
0.998	1.002

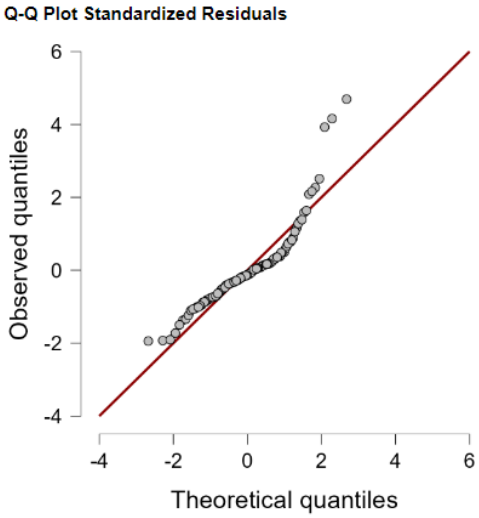
According to the condition index, there are no variables greater than 30 and has Tolerance level close to 1 and VIF close to 1 hence the model passed the multicollinearity test.

Summary:

Model Diagnostics/Assumptions	**Results**
Normality (Histogram/Formal Test)	**Failed**
Homoscedasticity (Variance Plot)	**Failed**
Linearity (Plots)	**Failed**
Independence (DW Test)	**Failed**
Multicollinearity (Collinearity Diagnostics	**Passed**

Normality Test (Backward Pass)





Model Summary - Total (Hg)

Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC	Durbin-Watson		
							Autocorrelation	Statistic	p
M_0	0.521	0.271	0.237	2.566	646.404	669.646	0.191	1.616	0.016
M_1	0.519	0.269	0.241	2.560	644.774	665.111	0.196	1.607	0.016
M_2	0.513	0.263	0.240	2.561	643.976	661.408	0.192	1.614	0.019
M_3	0.505	0.255	0.238	2.566	643.452	657.979	0.243	1.513	0.003
M_4	0.500	0.250	0.238	2.564	642.349	653.970	0.263	1.473	0.002

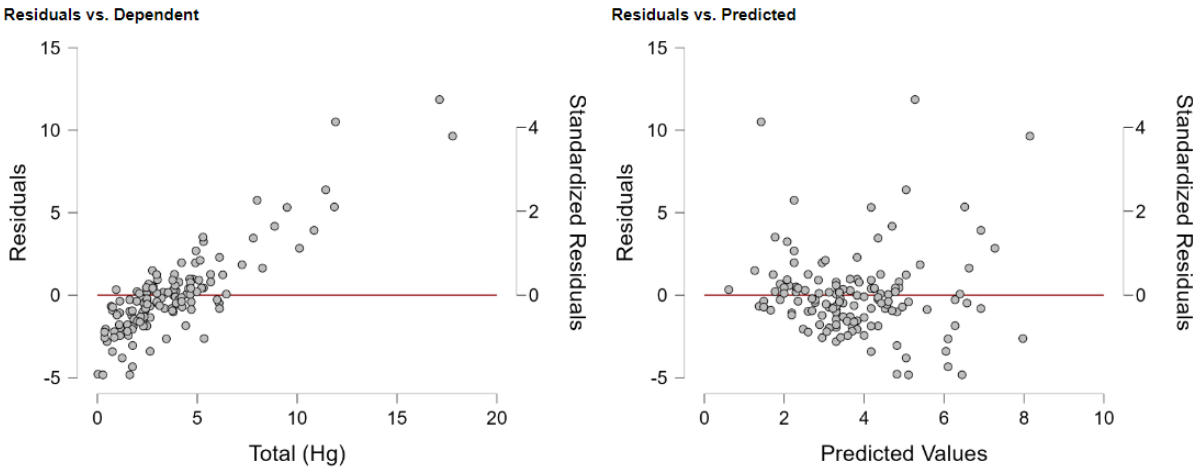
Descriptive Statistics

Descriptive Statistics

	Total (Hg)
Std. Deviation	2.939
Shapiro-Wilk	0.816
P-value of Shapiro-Wilk	< .001

The model failed in Normality test the Shapiro-wilk P value is less than .001 which means we'll reject the null hypothesis that residuals follow normal distribution.

Homoscedasticity



Upon visual assessment, there is a change in the spread or variance of the data points, hence the model does not satisfy homoscedasticity

Linearity

The plot Residuals vs. Predicted does not show linearity thus it failed the test. This could mean that we might need to log transform our model to fully capture the relationship between dependent and independent variables.

Independence

Model Summary - Total (Hg)

							Durbin-Watson		
Model	R	R^2	Adjusted R^2	RMSE	AIC	BIC	Autocorrelation	Statistic	p
M_0	0.521	0.271	0.237	2.566	646.404	669.646	0.191	1.616	0.016
M_1	0.519	0.269	0.241	2.560	644.774	665.111	0.196	1.607	0.016
M_2	0.513	0.263	0.240	2.561	643.976	661.408	0.192	1.614	0.019
M_3	0.505	0.255	0.238	2.566	643.452	657.979	0.243	1.513	0.003
M_4	0.500	0.250	0.238	2.564	642.349	653.970	0.263	1.473	0.002

Based from Durbin-Watson test, the DW statistic of 1.473, suggests positive auto correlation in residuals which means they are not independent.

Multicollinearity

Collinearity Diagnostics ▼

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions						
				(Intercept)	age	restime (hours)	height (cm)	weight (kg)	number of fish meals per week	owned_boat
M_0	1	5.632	1.000	0.000	0.001	0.007	0.000	0.000	0.008	0.009
	2	0.554	3.188	0.000	0.001	0.147	0.000	0.000	0.009	0.746
	3	0.454	3.524	0.000	0.001	0.494	0.000	0.001	0.007	0.206
	4	0.328	4.146	0.000	0.002	0.013	0.000	0.001	0.947	0.002
	5	0.027	14.466	0.002	0.951	0.301	0.003	0.028	0.026	0.036
	6	0.005	34.248	0.062	0.043	0.037	0.045	0.956	0.001	0.001
	7	7.416×10^{-4}	87.147	0.935	0.002	0.001	0.952	0.014	0.002	0.001
M_1	1	5.167	1.000	0.000	0.001	0.008	0.000	0.000	0.010	
	2	0.472	3.308	0.000	0.000	0.647	0.000	0.001	0.001	
	3	0.328	3.970	0.000	0.002	0.014	0.000	0.001	0.965	
	4	0.028	13.616	0.002	0.954	0.293	0.002	0.026	0.021	
	5	0.005	32.788	0.062	0.041	0.036	0.044	0.958	0.001	
	6	7.420×10^{-4}	83.446	0.935	0.001	0.002	0.953	0.014	0.003	
M_2	1	4.216	1.000	0.000	0.002	0.013		0.000	0.015	
	2	0.445	3.079	0.001	0.000	0.678		0.001	0.034	
	3	0.309	3.692	0.002	0.005	0.001		0.002	0.934	
	4	0.026	12.702	0.022	0.928	0.269		0.057	0.016	
	5	0.004	33.067	0.975	0.065	0.040		0.939	0.000	
M_3	1	3.651	1.000	0.001	0.004			0.001	0.021	
	2	0.309	3.436	0.002	0.007			0.002	0.945	
	3	0.036	10.133	0.025	0.963			0.042	0.033	
	4	0.004	30.187	0.972	0.027			0.955	0.001	
M_4	1	2.703	1.000	0.001				0.001	0.043	
	2	0.292	3.040	0.004				0.004	0.957	
	3	0.004	25.667	0.995				0.995	0.000	

Collinearity Statistics	
Tolerance	VIF
0.621	1.610
0.638	1.569
0.897	1.115
0.891	1.122
0.920	1.087
0.968	1.033
0.636	1.572
0.647	1.547
0.897	1.115
0.892	1.122
0.929	1.077
0.638	1.568
0.653	1.532
0.988	1.012
0.931	1.074
0.933	1.072
0.997	1.003
0.933	1.071
0.998	1.002
0.998	1.002

According to the condition index, on the last model utilized, there are no variables greater than 30 with Tolerance and VIF close to 1, hence the model passed the multicollinearity test.

Summary:

Model Diagnostics/Assumptions	**Results**
Normality (Histogram/Formal Test)	**Failed**
Homoscedasticity (Variance Plot)	**Failed**
Linearity (Plots)	**Failed**
Independence (DW Test)	**Failed**
Multicollinearity (Collinearity Diagnostics	**Passed**

The Full model did not pass any test while both forward and backward pass only succeed on multicollinearity test which agrees with my answer in number 3 that both forward and backward pass models has the same robustness compared to Full model LR.