Creating and Interpreting Multiple Linear Regression

Variables for Analysis & Assumptions

This analysis aims to explore the potential correlation between body weight, chest girth, age, and body length in Florida Black Bears. We will be attempting to use the explanatory variables (chest girth, age, and body length) to predict the response variable (body weight). It will involve distinct examinations for both male and female Black Bears.

	sex	age	body_wt	chest_girth	body_len
3	MALE	1.50	32.21	62.0	131.0
4	MALE	1.75	36.29	63.0	121.0
6	MALE	1.67	35.38	63.5	112.0
7	MALE	1.75	43.09	66.0	139.0
11	MALE	1.58	44.45	67.0	141.0

	sex	age	body_wt	chest_girth	body_len
0	FEMALE	1.67	26.31	54.2	113.3
1	FEMALE	1.42	27.22	59.0	127.0
2	FEMALE	1.67	22.68	60.0	109.0
5	FEMALE	2.67	44.45	63.0	123.0
8	FEMALE	1.92	40.82	66.0	136.0

Assumptions:

1. Linearity or Mean of Zero: The mean of each residual for each set of values for the predictor variables is zero. Equivalently, this assumption says that the response variable is a linear function of each of the predictor variables.[1]
2. Independence: Observations are independent. / The residuals are independent; no clear patterns exist. [1]
3. Normality: The residuals of each set of values for the predictor variables form a normal distribution. [1]
4. Homoscedasticity or Constant variance: For each value of the predictor (x), the probability distribution of the regression error has constant equal variance. [1]

Regression Models:

Model 1: Body Weight vs. Chest Girth

Recall from previous SLR analysis of the model had the below aspects:

Liner Regression Model 1 (Body Weight vs. Chest Girth)			
Male Bears		Female Bears	
r	0.9812	r	0.9611
R^2	0.96	R^2	0.92
Equation	$\hat{y}$ = -108.98 + 2.12x	Equation	$\hat{y}$ = -81.82 + 1.77x

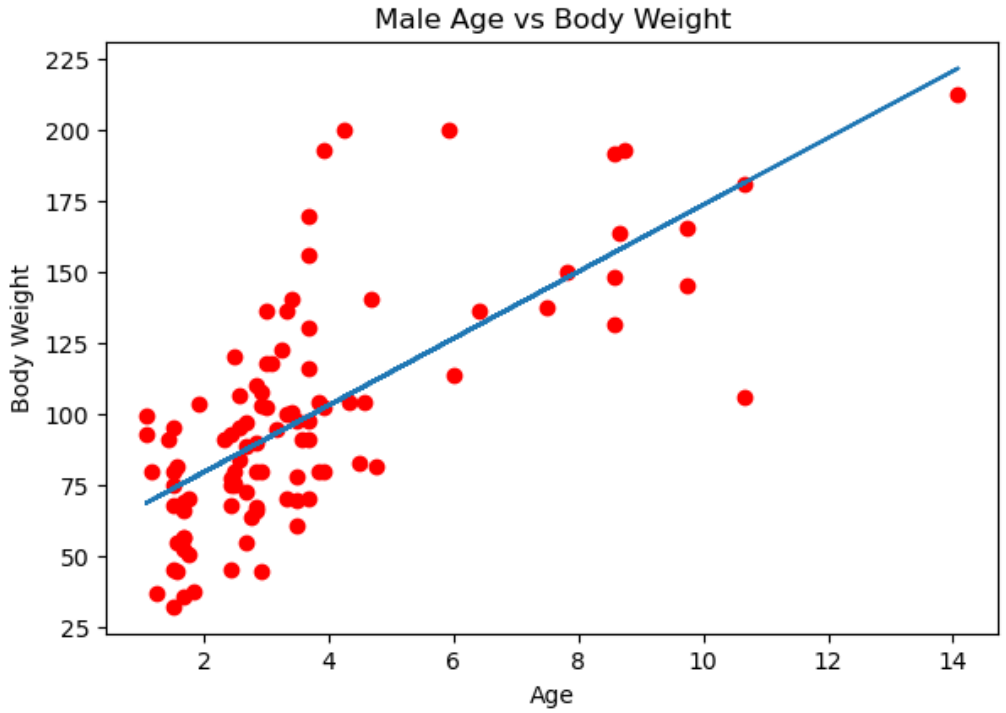
In this regression equation, " $\hat{y}$ " denotes the dependent variable “Body Weight” under investigation, which we aim to predict or understand. "x," on the other hand, signifies the independent variable, “Chest Girth” employed to make predictions regarding the dependent variable. In this context, "b_0" serves as the y-intercept, representing the value of " $\hat{y}$ " or “Body Weight” when "x" or ‘Chest Girth’ equals zero. It essentially defines the baseline value of the dependent variable when the independent variable exerts no influence. "b_1," meanwhile, denotes the slope of the regression coefficient and quantifies the change in the dependent variable "Body Weight" for each one-unit alteration in the independent variable "Chest Girth."

Model 2: Body Weight vs. Chest Girth and Age

This model’s equations will be in form: $\hat{y} = b_0 + b_1x_1 + b_2x_2$

In this regression equation, " $\hat{y}$ " denotes the dependent variable ‘Body Weight’ under investigation, which we aim to predict or understand. " x_1" represents the independent variable “Chest Girth” and “x_2” represents the independent variable “Age”. Both variables are being used to make predictions regarding the dependent variable “Body Weight”. In this context, "b_0" still serves as the y-intercept, representing the value of " $\hat{y}$ " or “Body Weight’ when " x_1" or “Chest Girth” and " x_2" or “Age” equal zero. It essentially defines the baseline value of the dependent variable when the independent variables exert no influence. "b_1" denotes the slope of the regression coefficient and quantifies the change in the dependent variable "Body Weight" for each one-unit alteration in the independent variable "Chest Girth” and "b_2" represents the slope of the regression coefficient and quantifies the change in the dependent variable "Body Weight" for each one-unit alteration in the independent variable "Age”.

Male Bear Multiple Liner Regression Equation: $\hat{y} = -98.47 + 1.93x_1 + 2.30x_2$



The scatter plot illustrating the correlation between body weight and age in male bears reveals a relatively weak positive relationship. This is apparent from the upward slope of the regression line, signifying a positive trend in the relationship. However, the strength of this relationship is best observed by examining the data points. These points display a loose clustering towards the lower end of the regression line, with fewer and more scattered data points along the middle and upper portions.

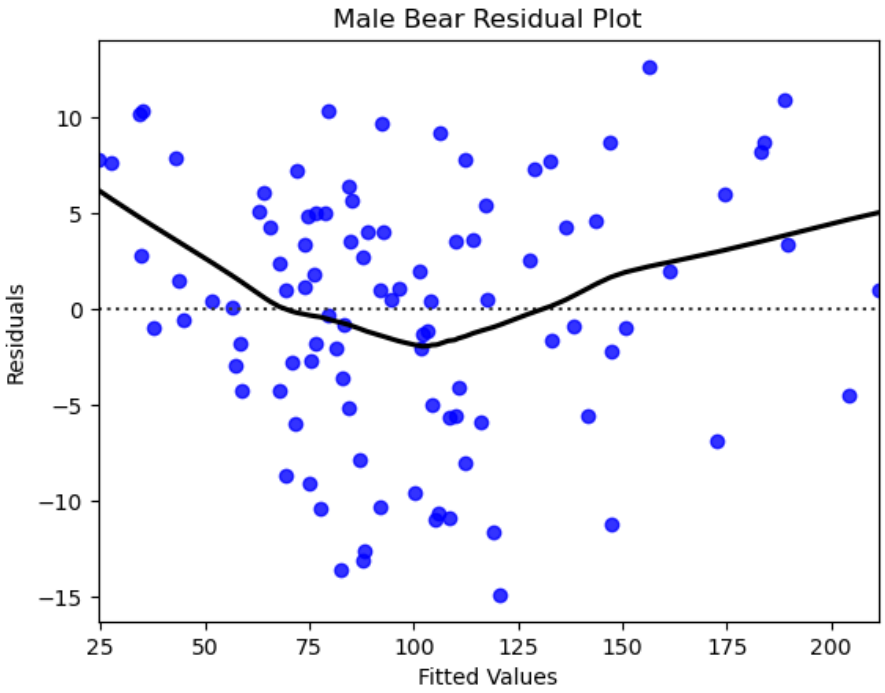
It's worth noting that a significant number of data points are concentrated in the lower region of the scatter plot, indicating that many male bears in the dataset share similar body weight and age values.

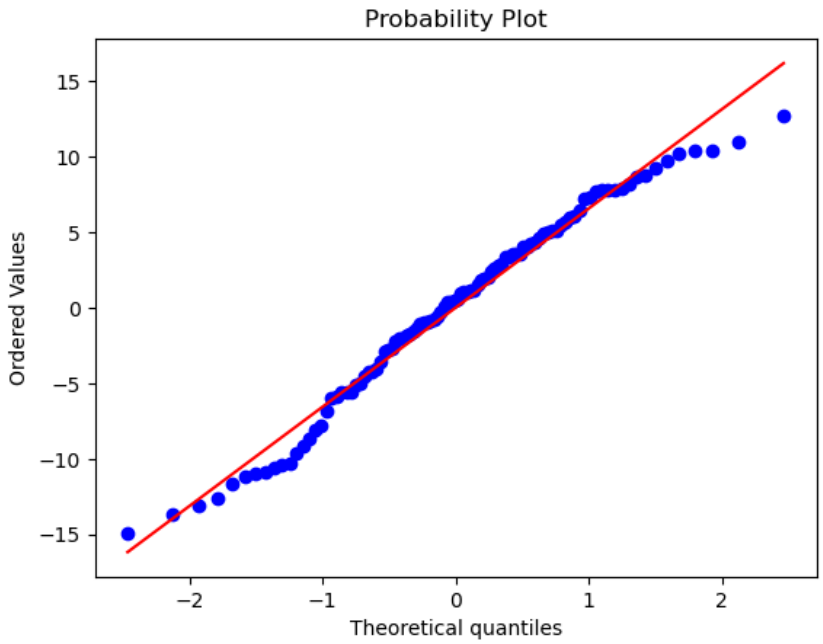
The strength of the relationship can be more accurately assessed through the correlation coefficient "r." The magnitude of the correlation coefficient, denoted as |r|, serves as an indicator of the relationship's strength. When |r| falls between 0 and 0.40, it is generally regarded as weak, between 0.40 and 0.80 as moderate, and between 0.80 and 1 as strong. In the specific case of the relationship between male bear body weight and age, the calculated value of r is 0.71, signifying a moderate relationship. To gain a more comprehensive understanding of the nature and strength of this relationship, a more in-depth analysis would be necessary.

Male Bear Coefficient of Determination R^2 : 0.974

R^2_{adj}: 0.974

Based on the R^2 value, it is apparent that 97.4% of the variability in body weight can be accounted for by the regression model. This notably high R^2 value indicates that the model effectively explains a significant portion of the variance in the data, underscoring its reliability in portraying the relationship between the variables. Nonetheless, it is crucial to note that while this model is informative, it may not necessarily represent the best-fit model. Therefore, further evaluation of the relationship and the model is advisable for a more comprehensive understanding.



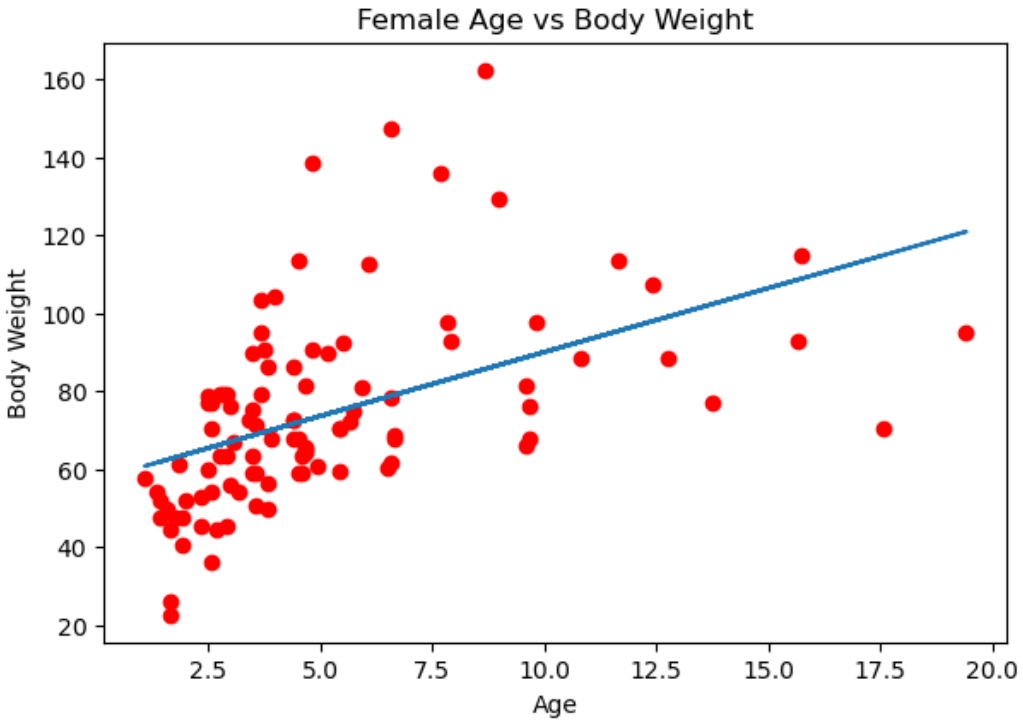


Upon examination of the residual plot, a subtle funnel pattern appears to be present. This observation could potentially suggest a departure from the assumptions of linearity or residuals having a mean of zero. It's worth noting, though, that this pattern is rather slight, and its impact on the model's validity may be minimal. To ascertain the true extent of this effect, further investigation is advisable. It's worth mentioning that no other conspicuous patterns are evident in the residual plot. Additionally, the QQ plot suggests that the data follows a relatively normal distribution.

Model Prediction using chest girth of 66cm and age of 7.5:

The estimated weight of a male black bear with a chest girth of 66 cm and an age of 7.5 is approximately 46.16 kg. Upon closer examination, it is essential to scrutinize the validity of this predicted value. When considering the data individually, the mean weight of a male bear aged between 7 and 8.5 years is 144.09 kg, and the mean weight of a male bear with a chest girth ranging from 60 to 70 cm is 38.74 kg. This leads to the observation that this model appears to be significantly influenced by chest girth and may not provide accurate predictions for age-related data.

Female Bear Multiple Liner Regression Equation: $\hat{y} = -78.34 + 1.69x_1 + 0.71x_2$



In a manner similar to the male bear data, the scatter plot depicting the correlation between body weight and age in female bears also demonstrates a relatively weak positive relationship. This is evident from the upward slope of the regression line, indicating a positive trend in the relationship. However, the strength of this relationship is best discerned by examining the data points, which exhibit a loose clustering towards the lower end of the regression line, with fewer and more scattered data points along the middle and upper segments.

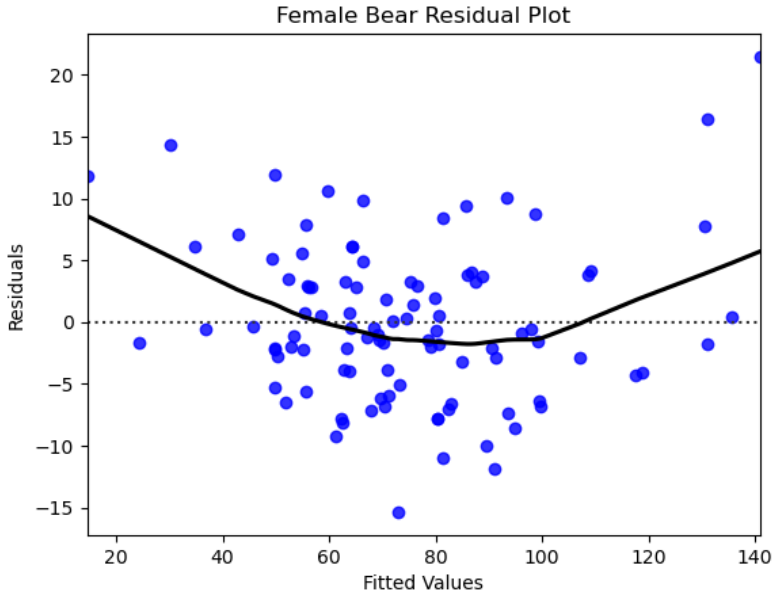
Furthermore, much like the male bear dataset, a noteworthy concentration of data points is observed in the lower region of the scatter plot, suggesting that numerous female bears within the dataset share similar body weight and age values.

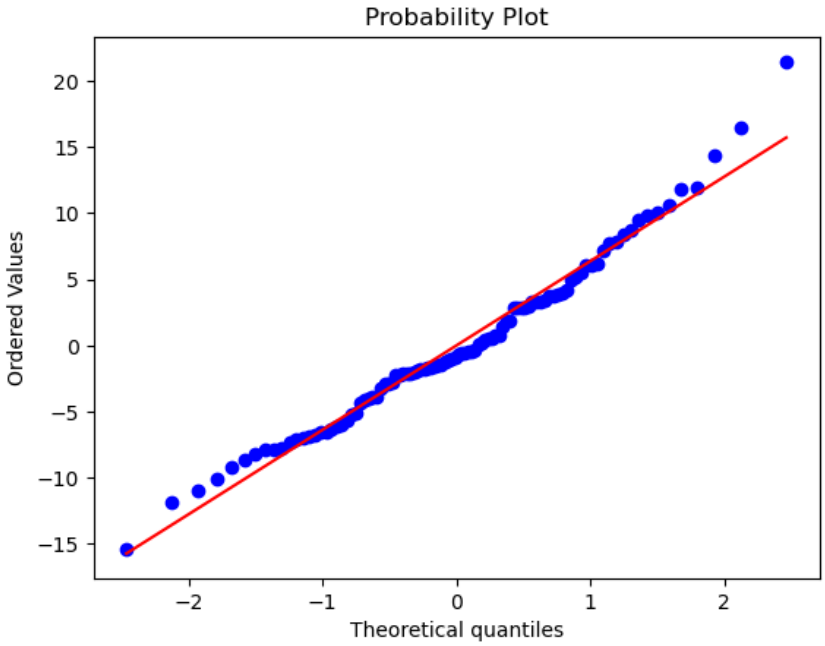
The strength of the relationship can be better assessed by examining the correlation coefficient, denoted as "r." In the context of female bear body weight versus age, the computed value of r is 0.49. This result still indicates a moderate correlation, although notably lower when compared to the male bear data. For a more comprehensive understanding of the nature and strength of this relationship, a more in-depth and rigorous analysis would be necessary.

Female Bear Coefficient of Determination R^2 : 0.933

R^2_{adj}: 0.931

Based on the R^2 value, it is apparent that 93.3% of the variability in body weight can be accounted for by the regression model. This notably high R^2 value indicates that the model effectively explains a significant portion of the variance in the data, underscoring its reliability in portraying the relationship between the variables. Nonetheless, it is crucial to note that while this model is informative, it may not necessarily represent the best-fit model. Therefore, further evaluation of the relationship and the model is advisable for a more comprehensive understanding.





Upon examining the residual plot pertaining to the female dataset, a slightly more pronounced funnel pattern becomes apparent. This observation raises the possibility of a departure from the assumptions of linearity or residuals with a mean of zero. To comprehensively assess the effect of this pattern on the model's validity, further investigation is warranted. It's worth noting that no other conspicuous patterns are discernible in the residual plot. Furthermore, the QQ plot indicates that the data adheres to a relatively normal distribution.

Model Prediction using chest girth of 66cm and age of 7.5:

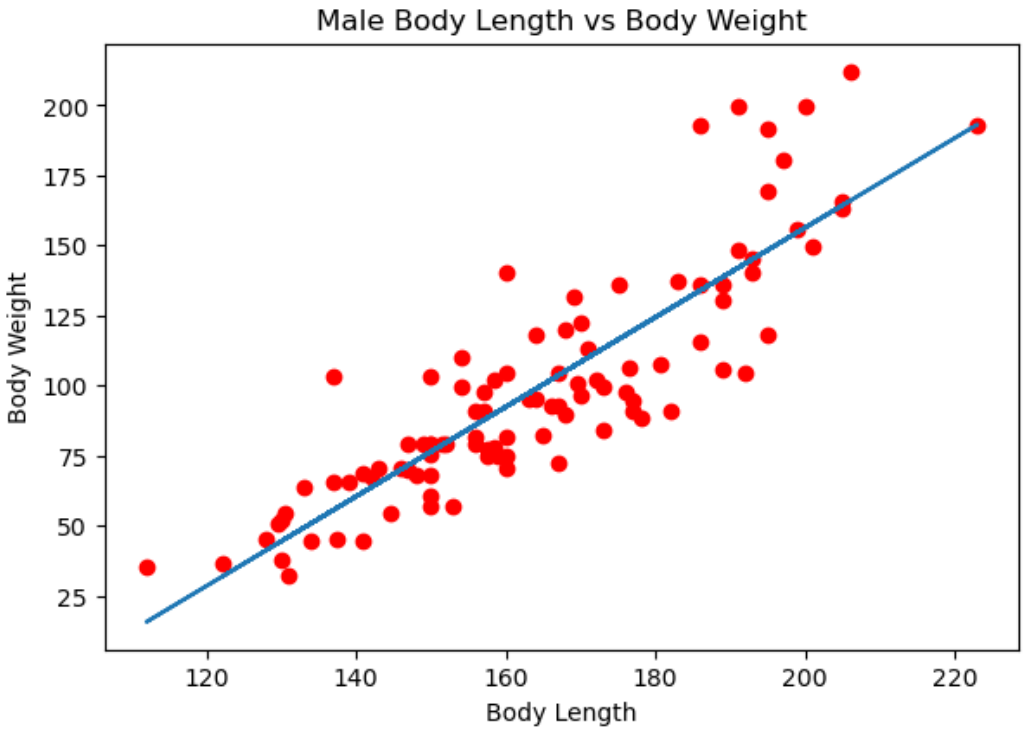
The expected weight of a female black bear with a chest girth of 66cm and an age of 7.5 is 38.52 kg. The estimated weight of a female black bear with a chest girth of 66 cm and an age of 7.5 is approximately 46.16 kg. Upon closer examination, it is essential to scrutinize the validity of this predicted value. When considering the data individually, the mean weight of a female bear aged between 7 and 8.5 years is 108.86 kg, and the mean weight of a female bear with a chest girth ranging from 60 to 70 cm is 38.83 kg. This leads to the observation that this model appears to be significantly influenced by chest girth and may not provide accurate predictions for age-related data.

Model 3: Body Weight vs. Chest Girth, Age, and Body Length

This model's equations will be in form: $\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_2x_2$

In this regression equation, " $\hat{y}$ " denotes the dependent variable 'Body Weight' under investigation, which we aim to predict or understand. " x_1" represents the independent variable "Chest Girth" and "x_2" represents the independent variable "Age". Both variables are being used to make predictions regarding the dependent variable "Body Weight". In this context, "b_0" still serves as the y-intercept, representing the value of " $\hat{y}$ " or "Body Weight' when " x_1" or "Chest Girth" and " x_2" or "Age" equal zero. It essentially defines the baseline value of the dependent variable when the independent variables exert no influence. "b_1" denotes the slope of the regression coefficient and quantifies the change in the dependent variable "Body Weight" for each one-unit alteration in the independent variable "Chest Girth", "b_2" represents the slope of the regression coefficient and quantifies the change in the dependent variable "Body Weight" for each one-unit alteration in the independent variable "Age" and "b_3" represents the slope of the regression coefficient and quantifies the change in the dependent variable "Body Weight" for each one-unit alteration in the independent variable "Body Length".

Male Bear Multiple Liner Regression Equation: $\hat{y} = -108.08 + 1.81x_1 + 2.10x_2 + 0.13x_3$



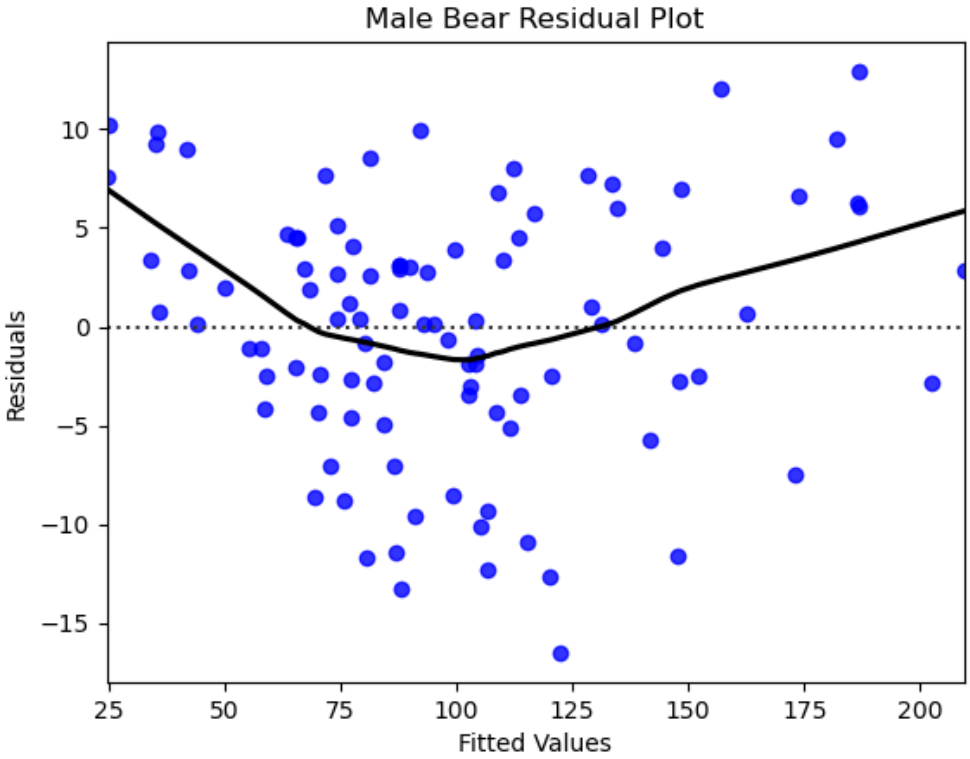
The scatter plot depicting the relationship between body weight and body length in male bears unveils a notably robust positive correlation. This is evident through the upward inclination of the regression line, signifying a pronounced positive trend in this relationship. While the data points are relatively evenly dispersed along the regression line, there is a slight degree of scattering.

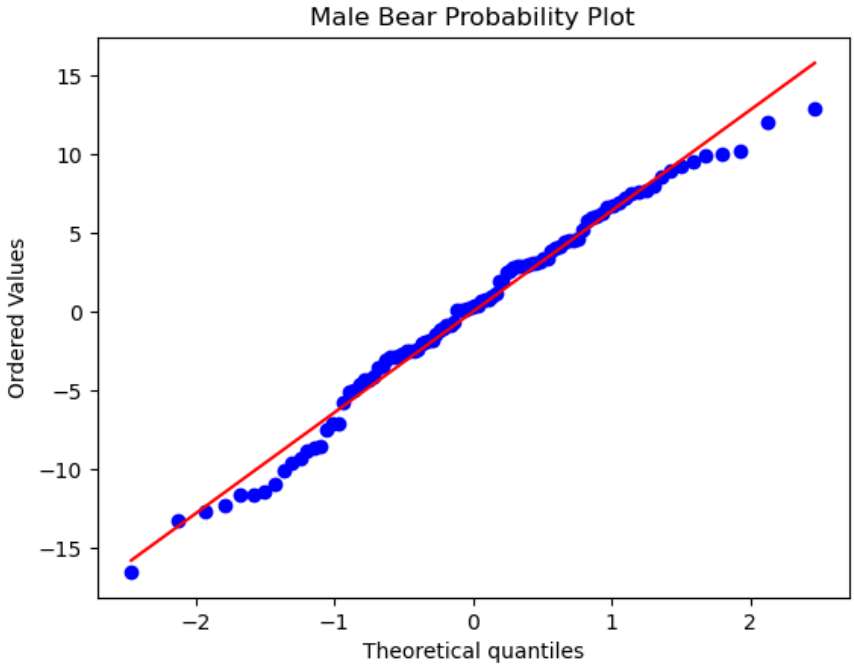
A more precise assessment of the strength of this correlation can be achieved by scrutinizing the correlation coefficient, commonly represented as "r." In the context of the male bear's body weight versus body length, the computed value of "r" stands at 0.87. This outcome implies a strong correlation.

Male Bear Coefficient of Determination R^2 : 0.976

R^2_{adj}: 0.975

Based on the R^2 value, it is apparent that 97.6% of the variability in body weight can be accounted for by the regression model. This notably high R^2 value indicates that the model effectively explains a significant portion of the variance in the data, underscoring its reliability in portraying the relationship between the variables. Nonetheless, it is crucial to note that while this model is informative, it may not necessarily represent the best-fit model. Therefore, further evaluation of the relationship and the model is advisable for a more comprehensive understanding.





Upon scrutinizing the residual plot, a funnel pattern seems to emerge. This observation might imply a deviation from the assumptions of linearity or residuals having a mean of zero. However, it's essential to note that the impact of this pattern on the model's validity may be relatively minor. To determine the precise extent of this effect, further investigation is recommended. It's noteworthy that no other striking patterns are discernible in the residual plot. Moreover, the QQ plot indicates that the data conforms to a relatively normal distribution.

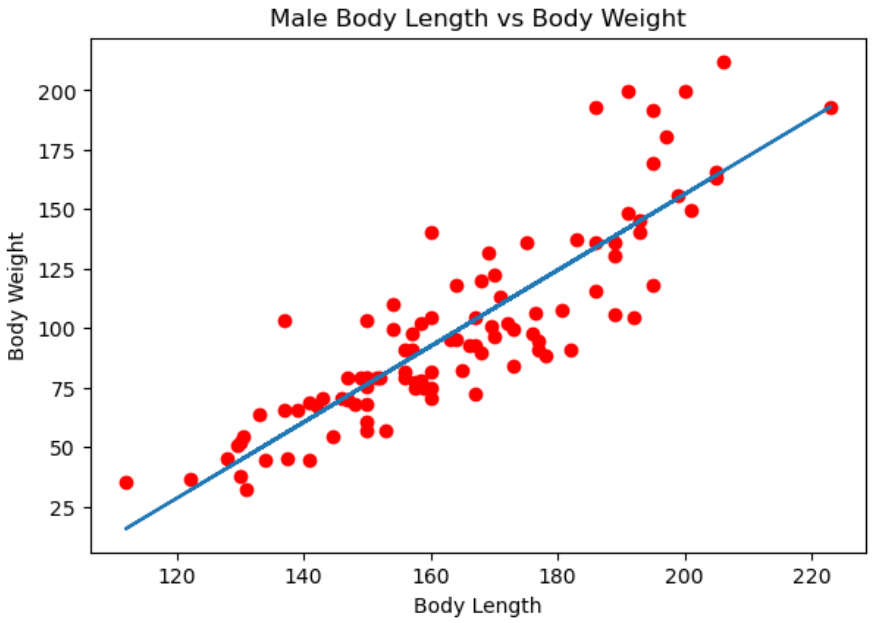
Confidence Interval for the mean expected/predicted weights: (27.419, 63.996)

Model Prediction using chest girth of 66 cm and age of 7.5 and body length is 140 cm:

The expected weight of a male black bear, given a chest girth of 66 cm, an age of 7.5 years, and a body length of 140 cm, is approximately 44.68 kg. While this forecast falls within the confidence interval of predictions, a more thorough examination of this predictive value is warranted to ensure its reliability.

Upon closer inspection of the data in isolation, it becomes evident that the mean weight of a male bear between the ages of 7 and 8.5 years is around 144.09 kg. Similarly, the average weight of a male bear with a chest girth falling within the range of 60 to 70 cm is approximately 38.74 kg, and for those with a body length ranging from 130 cm to 150 cm, the average weight is roughly 64.28 kg. This analysis indicates a significant influence of chest girth on the model's predictions and raises concerns about its accuracy when it comes to age-related data.

Female Bear Multiple Liner Regression Equation: $\hat{y} = -97.28 + 1.53x_1 + 0.58x_2 + 0.22x_3$



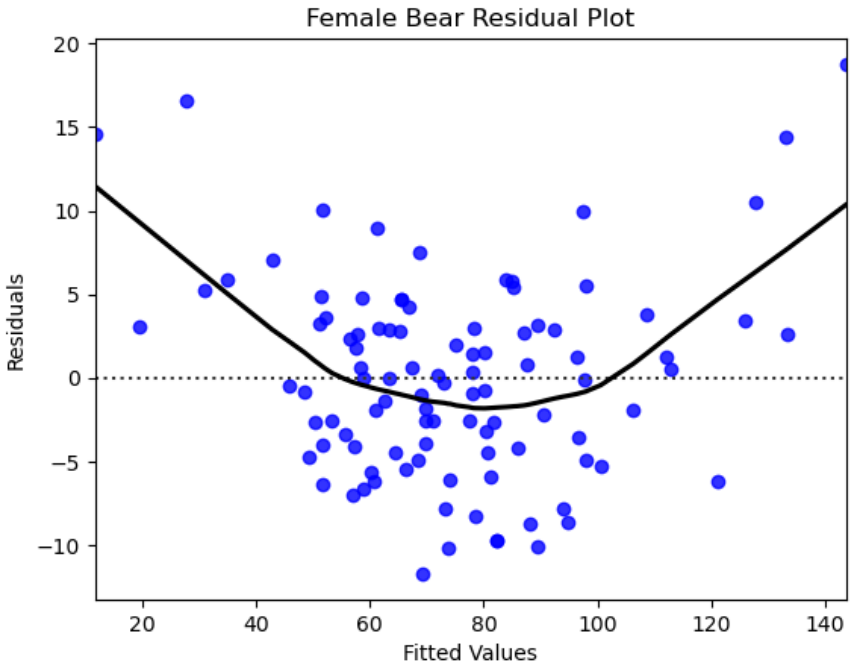
Similar to the male bear data, the scatter plot illustrating the correlation between body weight and body length in female bears reveals a moderately strong positive correlation. This is evident from the upward slope of the regression line, indicating a clear positive trend in this relationship. While the data points are relatively evenly distributed along the regression line, there is a noticeable difference in clustering, with a tighter grouping in the lower portion and a wider spread in the upper portion.

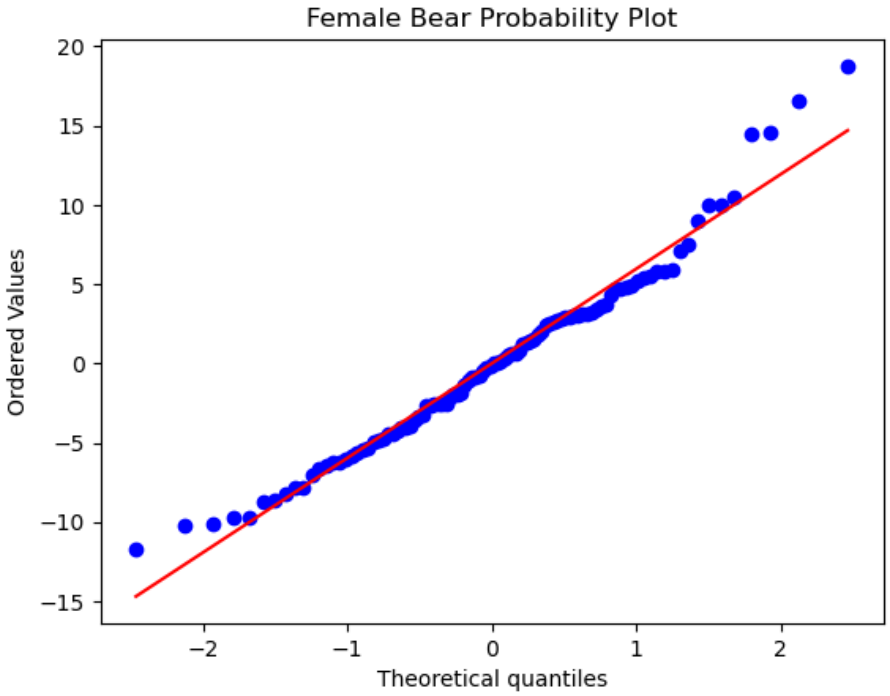
To obtain a more precise assessment of the strength of this correlation, it is advisable to examine the correlation coefficient, typically denoted as "r." In the context of female bears' body weight versus body length, the computed value of "r" is 0.77. This result suggests a moderate correlation.

Female Bear Coefficient of Determination R^2 : 0.941

R^2_{adj}: 0.939

Based on the R^2 value, it is apparent that 94.1% of the variability in body weight can be accounted for by the regression model. This notably high R^2 value indicates that the model effectively explains a significant portion of the variance in the data, underscoring its reliability in portraying the relationship between the variables. Nonetheless, it is crucial to note that while this model is informative, it may not necessarily represent the best-fit model. Therefore, further evaluation of the relationship and the model is advisable for a more comprehensive understanding.





Upon a thorough analysis of the residual plot for the female dataset, a distinct funnel pattern becomes notably pronounced. This observation prompts consideration of a potential departure from the assumptions of linearity or residuals with a mean of zero. To comprehensively evaluate the impact of this pattern on the model's validity, further investigation is highly recommended. It is important to highlight that no other discernible patterns are evident in the residual plot. Additionally, the QQ plot provides evidence that the data conforms to a relatively normal distribution.

Confidence Interval for the mean expected/predicted weights: (22.685, 55.445)

Model Prediction using chest girth of 66 cm and age of 7.5 and body length is 140 cm:

The estimated weight of a female black bear, considering a chest girth of 66 cm, an age of 7.5 years, and a body length of 140 cm, is expected to be around 37.75 kg. However, it's crucial to critically assess the accuracy of this prediction. While this estimate falls within the anticipated mean confidence interval, a detailed examination of the data reveals significant variations.

Specifically, when we break down the data into individual factors, we find that the average weight of a female bear aged between 7 and 8.5 years is approximately 108.86 kg. Similarly, the average weight of a female bear with a chest girth falling within the 60 to 70 cm range is about 38.83 kg, and for those with a body length between 130 cm and 150 cm, the average weight is roughly 61.35 kg. These findings suggest that the model's predictions are notably influenced by chest girth and may not provide accurate results when it comes to age-related data.

Summary

Regression Model Comparisons

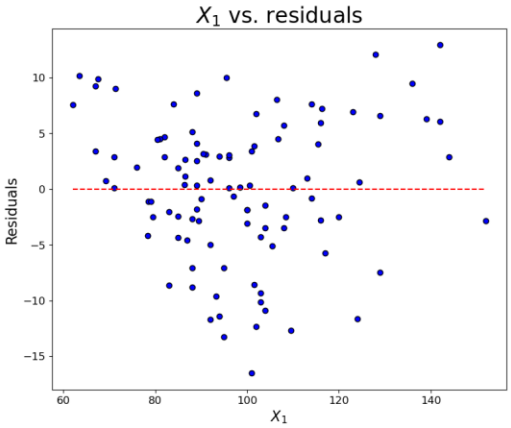
Male Bears

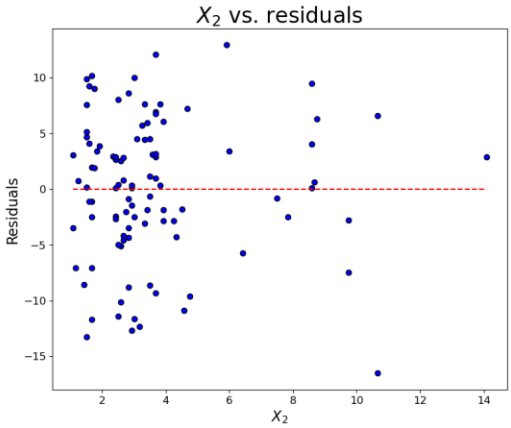
	r_{x1}	r_{x2}	r_{x3}	R^2	R^2_{adj}	VIF_{x1}	VIF_{x2}	VIF_{x3}	T-Test	F-Test Prob
Model 1	0.9812			0.963					0.000	
Model 2	0.9812	0.7114		0.974	0.974	1.6997	1.6997			9.12E-78
Model 3	0.9812	0.7114	0.8709	0.976	0.975	3.9289	1.8044	9.9539		3.48E-77

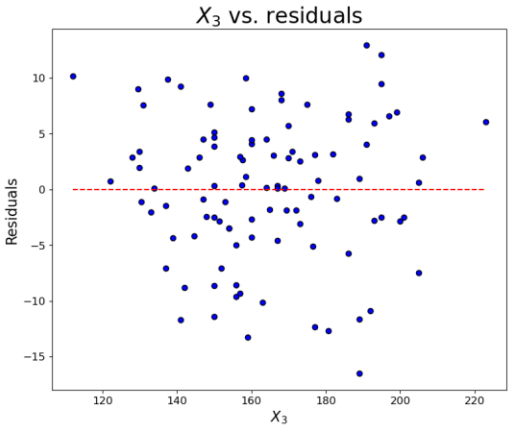
Female Bears

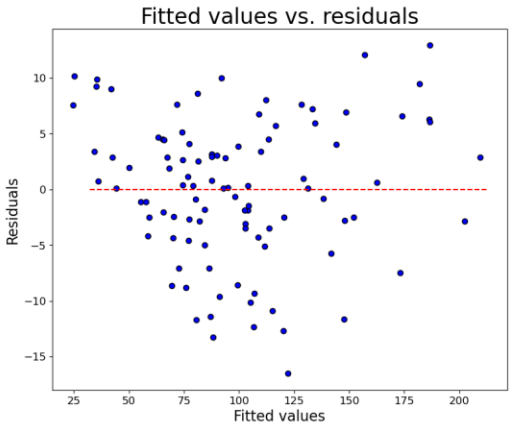
	r_{x1}	r_{x2}	r_{x3}	R^2	R^2_{adj}	VIF_{x1}	VIF_{x2}	VIF_{x3}	T-Test	F-Test Prob
Model 1	0.9611			0.924					0.000	
Model 2	0.9611	0.4888		0.933	0.931	1.2116	1.2116			1.29E-57
Model 3	0.9611	0.4888	0.7694	0.941	0.939	2.176	1.2532	2.1739		7.25E-59

Male Bear Residuals vs. Fitted values

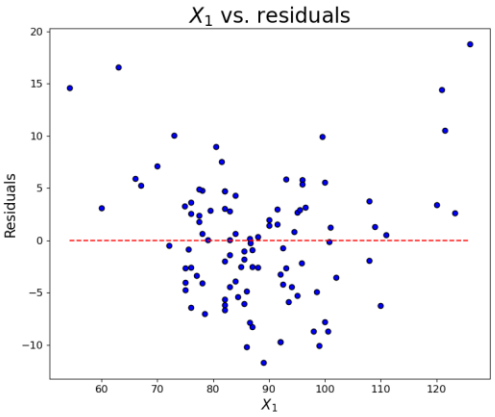


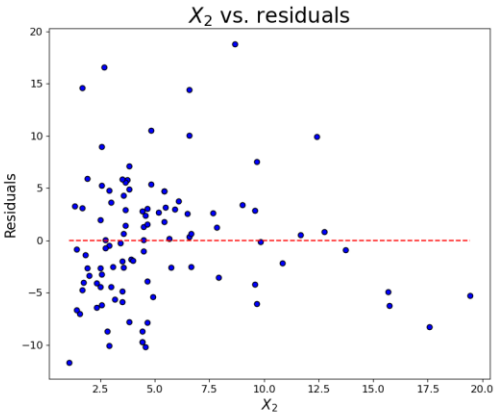


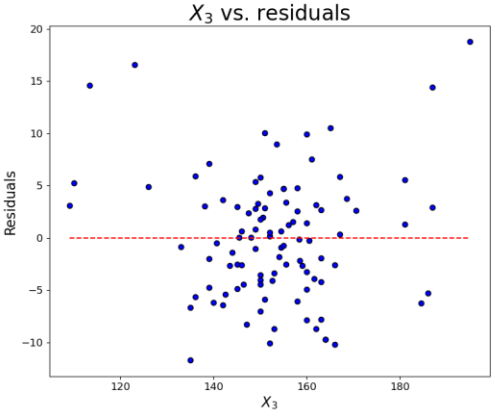


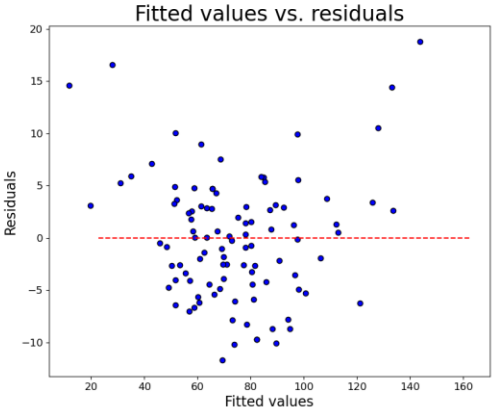


Female Bear Residuals vs. Fitted values









Upon thorough examination of all the models, my personal preference leans towards the original model, which relies on chest girth to predict body weight. Several analytical findings support this preference. Firstly, the visual correlation observed in the scatter plots appears to be the strongest within this particular model. Additionally, both the correlation coefficient (r) and the coefficient of determination (R2) exhibit greater strength in this model. It seems to be the most suitable fit for the sample data at hand.

Nonetheless, it's worth noting that while all the models hint at a robust linear relationship, as indicated by the scatter plots, high correlation coefficients, and substantial coefficients of determination, these observations alone do not validate the appropriateness of the regression model within this context. Several concerns have been identified with each model.

To gain a deeper understanding of the data and develop a more efficient model, further analysis is required. I believe that the dataset's most critical improvement lies in its organization, specifically in segregating the data by gender and further subdividing it into relevant age groups.

Once this restructuring is achieved, a more comprehensive analysis can be undertaken to identify the most suitable model, whether it's a Simple Linear Regression (SLR), Multiple Linear Regression, or a more complex regression model. By reevaluating the data, considering age group segmentation, and exploring alternative regression models, we can conduct a more precise and robust analysis that better aligns with the underlying characteristics of the dataset.