STAT 6210 Homework 3 Report

Xue Ming Wang (Vivian)
G20580112

Part A

The simulated dataset hw3a.csv has two variables, X and Y with 200 records. Part A will investigate the relationship between X and Y by exploring the joint density of (*X*, *Y*).

1. Correlation

1.1 Simple Statistics of Variable X and Y

Variable X		Variable Y	
Mean	0.1125	Mean	0.2045
Std Dev	1.0251	Std Dev	1.1383
Pearson Correlation: 0.76012			
Spearman Correlation: 0.7304			

Table 1. Simple statistics of variable X and Y

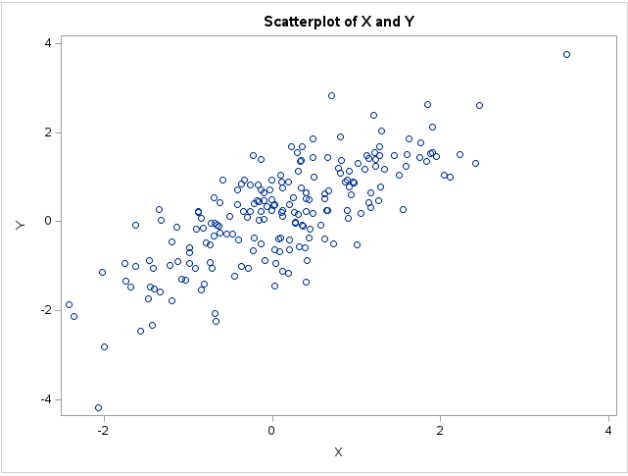

Figure 1. Scatterplot of X and Y

Table 1 indicated that the Pearson correlation is 0.76 and the Spearman correlation is 0.73, indicating a linear relationship between the two variables. This is also supported by the line pattern observed in Figure 1. Therefore, linear regression can be applied to analyze this dataset.

2. Joint Density of (x, y)

2.1 Simulate Joint Density for X and Y.

Variable X		Variable Y	
Mean	0.11	Mean	0.20
Variance	1.05	Variance	1.30
Standard Error	1.03	Standard Error	1.14

Table 2. simulate joint density for x and y.

2.2 Density Plot of X and Y

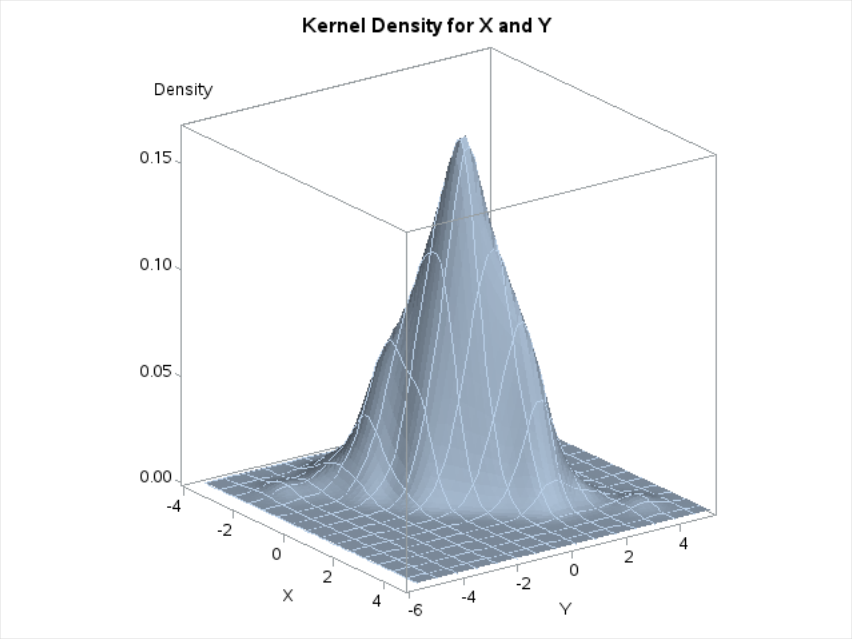

Figure 2. Density plot x and y

The simulated parameters are μ_x=0.11, μ_y=0.2, and σ_x=1.05, and σ_y=1.3 (Table 2). These parameters estimated the joint density distribution of variables X and Y, where figure 2 showed that visualizations of the density of X and Y.

2.3 Kernel Density Estimation with Different Bandwidths

In order to get a smoother estimate, we can adjust the bandwidth value and observe its effect on the density estimation. The 3 different bandwidths that will be used here are (0.5, 0.5), (1, 1), (2, 2).

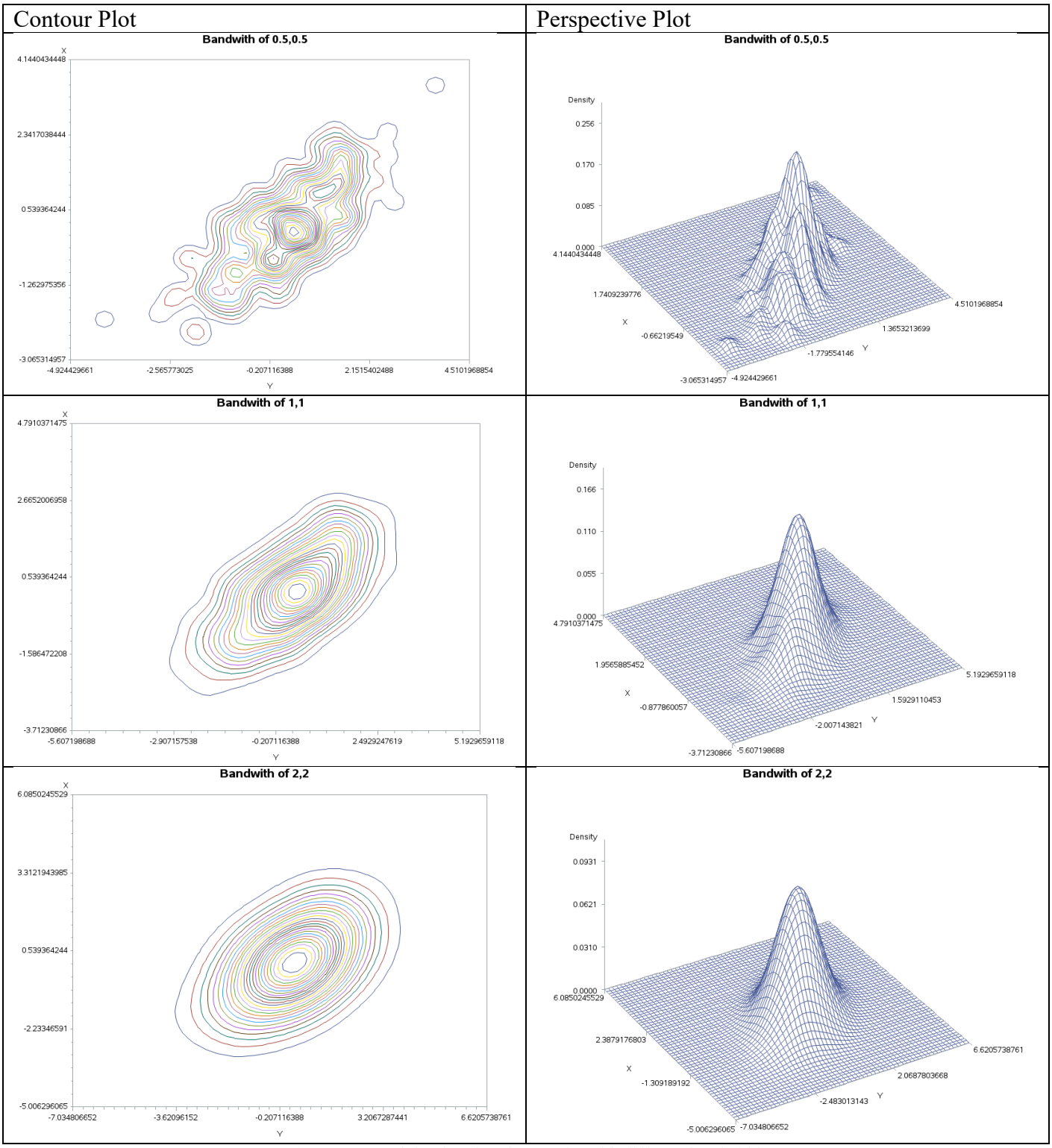

Figure 3. Kernel density plots for x and y with different bandwidth

Figure 3 displayed both contour plots (left side) and perspective plots (right side) with 3 different bandwidths. The bandwidth parameter plays a critical role in density estimation. A smaller bandwidth will result in a more variable estimate of the density function, which can capture more

detailed information about the distribution but may also introduce more noise. On the other hand, a larger bandwidth will result in a smoother estimate that is less sensitive to small variations in the data but may miss important features of the distribution. When the bandwidth was set to (0.5,0.5), the contour plot showed multiple centers and the 3D plot displayed more variability in the density surface. This suggests that the underlying distribution may not be well-approximated by a joint normal distribution. When the bandwidth was increased to (1,1), the contour plot showed a more symmetric and elliptical shape, which is a common feature of bivariate normal distributions. The lines in the contour plot were also wider than for the smaller bandwidth, indicating a smoother estimate. However, some details of the distribution may have been lost. Finally, when the bandwidth was set to (2,2), the contour plot showed an even smoother estimate with wider lines. This may have resulted in the loss of some of the finer details of the distribution.

3. Conclusion

Based on the analysis above, I think that the bandwidth of (1,1) could provide the best density estimate. A bandwidth of (0.5,0.5) is too specific for each data point, resulting in an estimator that is not generalizable to all points. When comparing to the joint density distribution (Figure 2), the contour plot with a bandwidth of (1,1) has a single center, which is more similar to the distribution. The bandwidth of (2,2) is too wide and results in a lack of detail. Therefore, it may not be the best choice for estimation. When the variables follow a normal distribution, it may be easier to analyze them in some cases. Therefore, a bandwidth of (1,1) is more suitable. Above all, the elliptical shape of the contour plot with a bandwidth of (1,1) suggests a bivariate normal distribution. The QQ-plots also suggest the same thing, where both variables x (Figure 4) and y (Figure 5) form a straight line, which means they are normally distributed. Since both x and y are approximately normally distributed, it is possible that the joint density of (X, Y) will have a bivariate normal distribution.

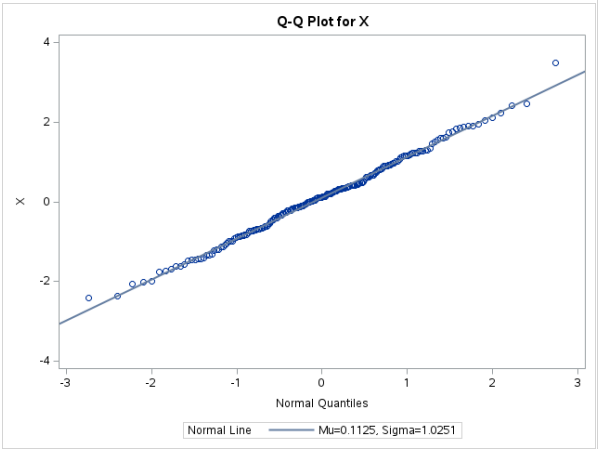

Figure 4. QQ plot of the normality of X.

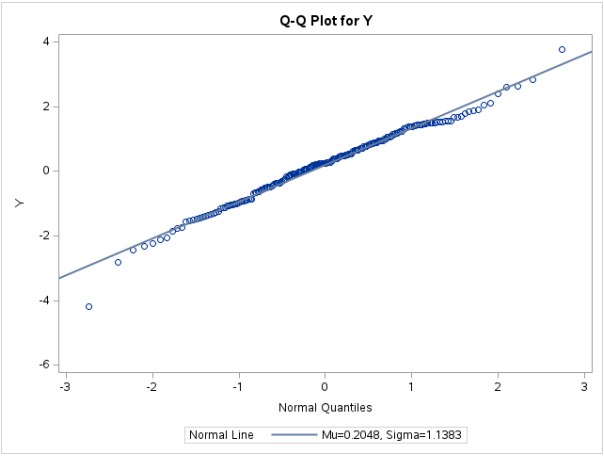

Figure 5. QQ plot of the normality of X

Part B

The dataset was obtained from a study of 237 children, which included the categorical variable sex and the continuous variables age, weight, and height. The aim was to investigate the relationship between weight and age under various conditions and develop a linear regression model for these two variables.

1. Correlation

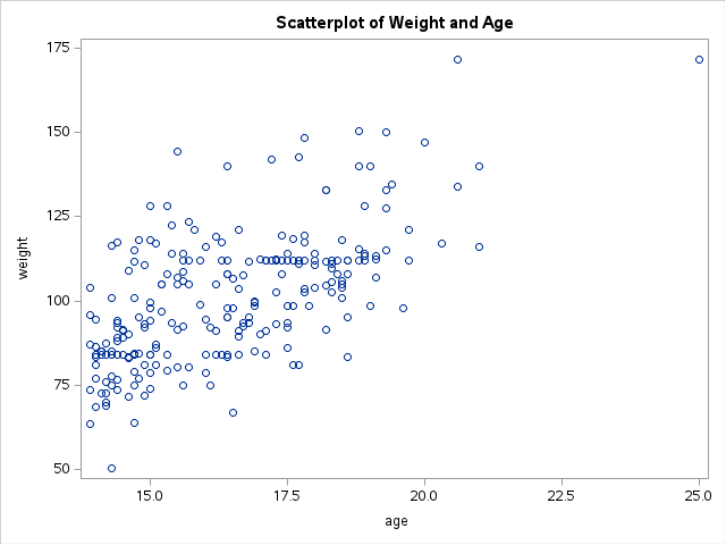

Figure 6 Scatterplot weight and age

Figure 6 indicated the relationship between weight and age with the scatterplot. As shown in the figure above, these two variables seem to have a linear relationship. To verify it, Pearson's correlation coefficient and its p-value will be conducted.

2. Pearson's Correlation Coefficient

Conditions	Pearson's correlation	P-value
Weight vs. Age	0.6346	<0.0001
Weight vs. Age, by Controlling Height	0.2743	<0.0001
Weight vs. Age, by Controlling Height, in Female	0.2361	0.0130
Weight vs. Age, by Controlling Height, in Male	0.3143	0.0004

Table 3. Pearson's correlation coefficients and their p-values

Table 3 showed the Pearson's correlation in different conditions. When only comparing weight with age, the Pearson's correlation coefficient was calculated to be 0.634, with a p-value of less than 0.0001, indicating these two variables a have linear relationship and there is a strong positive correlation between weight and age.

When comparing weight with age while controlling height, the Pearson's correlation coefficient was found to be 0.274 and the p-value was less than 0.0001. This suggests that there is a significant linear relationship between weight and age even after controlling for the effect of height.

When adding in different sex groups while controlling height, the Pearson's correlation coefficient for females is 0.2361 with a p-value of 0.0130, indicating a significant linear relationship between weight and age. Similarly, for males, the Pearson's correlation coefficient is 0.3143 with a p-value of 0.0004, indicating a significant linear relationship between weight and age while controlling for height. Weight and age have a linear relationship when height is controlled in both female and male sex groups. Under any condition, the results with Pearson's correlation suggest a significant linear relationship between weight and age.

3. Simple Linear Regression Between Weight and Age

Variable	Parameter Estimate	Standard Error	P-value
Intercept	−8.7935	8.8005	0.3187
Age	6.6959	0.5319	<0.0001

Table 4: Parameter estimates for simple linear regression

Source	Value
Model	P-value<0,0001
R-square	0.4002

Table 5: Analysis of variance for simple linear regression

The simple linear regression analysis between weight and age shows that the estimated model is weight = - 8.7935 + 6.6959 * age, and the p-value of age is less than 0.05, indicating a significant linear relationship between weight and age. However, the p-value of the intercept is greater than 0.05. The model is also significant since the analysis of variance also shows that the p-value of the model is less than 0.0001 (Table 5). Yet, the R-square value is only 0.4002 (Table 5), which means only approximately 40% of the variation in weight can be explained by age in the simple linear regression model. So, this simple regression model may not be a good fit for these two variables. A better model is needed.

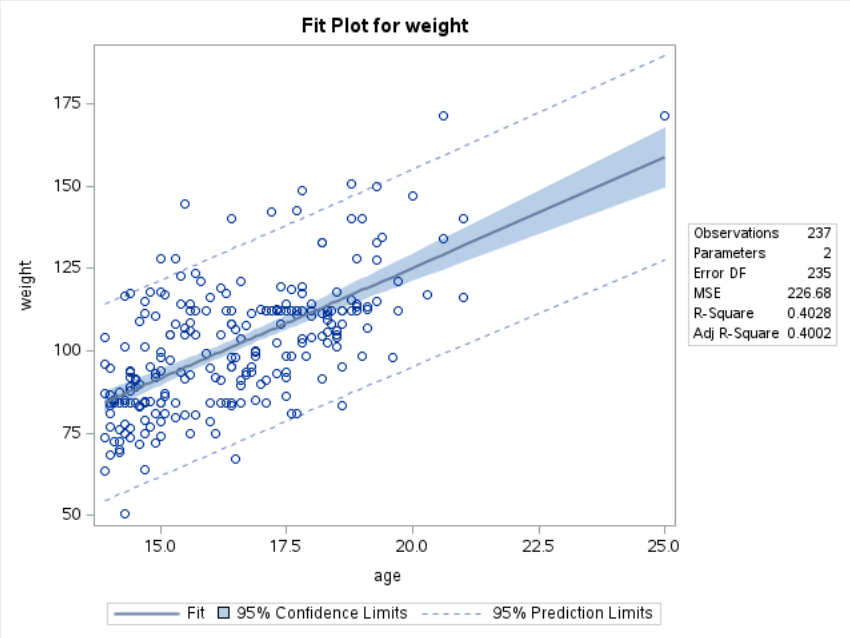

Figure 7. Fit plot for weight in simple linear regression

The plot in Figure 7 shows the fitted line along with the shaded area indicating the 95% confidence interval. However, there are still a significant number of data points outside this interval, indicating a substantial amount of variability in the data. So, a better model is needed.

4. Local Regression Between Weight and Age

Optimal Smoothing Criterion	Value
AICC	6.4194
Smoothing Parameter	0.4494

Table 6: Optimal Smoothing Criterion for local regression

Since the linear model may be too simple to achieve a satisfactory fit, we can use local regression to achieve a better fit. Table 6 shows the optimal smoothing criterion for local regression. The lower the AICC value, the better the fit of the model. In this case, the AICC value is 6.41943, which suggests a good fit for the model. The smoothing parameter is a parameter that controls the amount of smoothing in the local regression. A smaller value of the smoothing parameter produces a more flexible, wiggly curve that fits the data more closely, while a larger value produces a smoother curve that is less influenced by individual data points. Here, the smoothing parameter is 0.44937, which suggests a moderate amount of smoothing in the local regression.

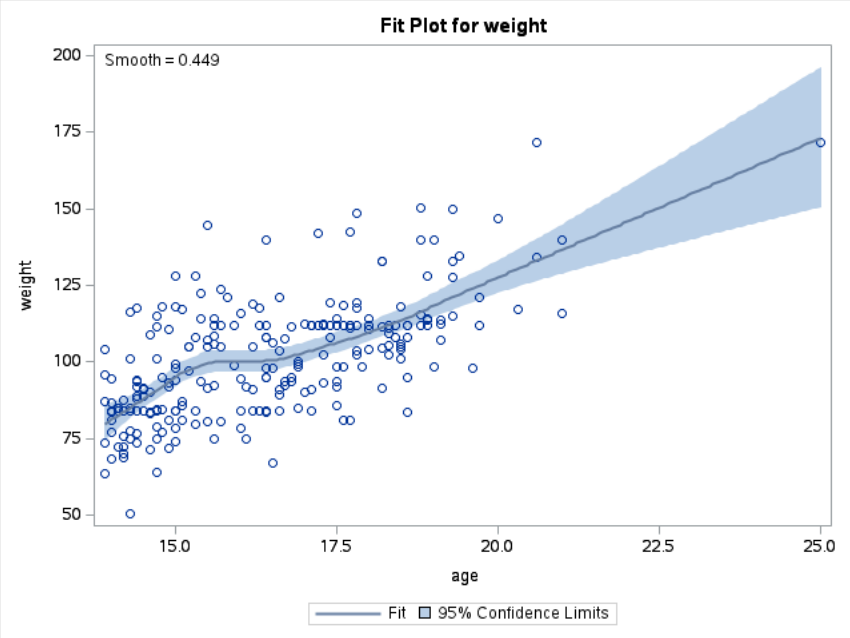

Figure 8. Fit plot for weight in local regression

Local regression involves fitting a regression surface to the data points within a selected neighborhood of xi to obtain a local approximation. Figure 8 displays the results of the local regression by examining the relationship between weight and age as well as the 95% interval of local regression. If the fitted regression line is a straight line, this indicates a linear relationship between the two variables. If the fitted line is curved, this indicates a non-linear relationship. In this plot, the majority of the fitted line appears to be straight, indicating that there is a linear relationship between weight and age. However, there are also some curved sections, indicating that the relationship between weight and age may not be strictly linear in all cases. This suggests that a local regression may be a better fit for this data than a simple linear regression.

5. Simple Linear Regression vs Local Regression

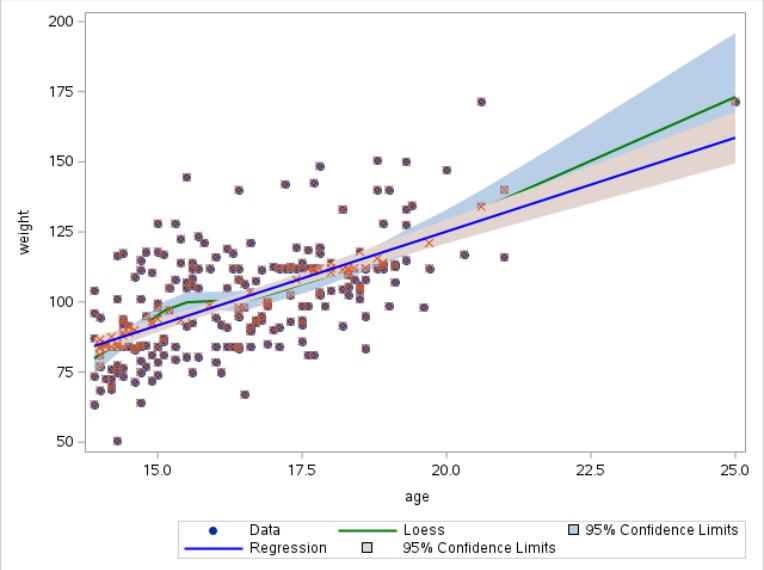

Figure 9. Simple regression vs local regression

While there are some differences between the confidence intervals of the simple and local regression models (Figure 9), they largely overlap. Taken together, the findings suggest that there is a linear relationship between age and weight.

6. Conclusion

The scatterplot of weight and age showed that the two variables have a linear relationship. The Pearson's correlation coefficient indicates that there is a correlation between weight and age, and this correlation persists even after accounting for controlled height and different sex groups. Moreover, after examining fitted plots for each model, the local regression model performs better than the simple regression model. The confidence intervals of the simple regression and local regression models show some differences, but they mostly overlap. Overall, all results suggest that weight and age follow a linear regression relationship.