Homework 3

Jiaying Liu
G45268292

February 27, 2018

1 Part A

The dataset hw3a.csv contains two variables x and y. We want to explore the relation between them. In a given dataset, we are often first interested in observing their relation.

1.1 correlation of two variable

Table 1. Basic statistical summary of variable x and y.

x		y	
Mean	0,112 45	Mean	0,204 84
std DEV	1,035 06	std Dev	1,381 28
Pearson Correlation	0,760 12	Spearman Correlation	0,730 36

Figure 1. Scatter of variable x and y

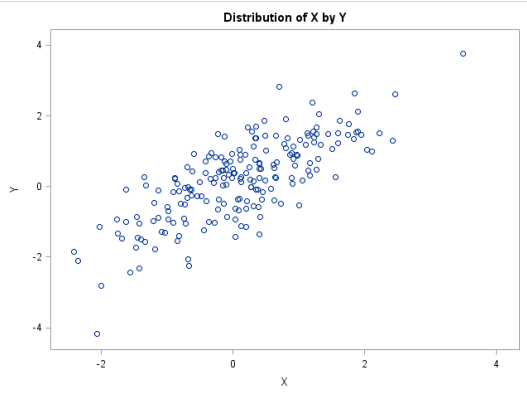


Table 1 show the basic statistical moment of two variables. Figure 1 show the Scatter Plot of two variable. From this two results, we can find that the mean of x is 0.11245 while the mean of y is 0.20484. We can find that the Pearson Correlation is 0.76 and the Spearman Correlation is 0.73036.Both of this two correlation show that this two variable have linear relation. And in the mean time, from the figure 1 it also obviously that two variable follow a line patent. In conclusion, we can use linear regression method to solve the data.

1.2 Joint density of (x, y)

Table 2. simulate joint density for x and y.

x		y	
Mu1	0,11	Mu2	0,20
sigma1	1,05	sigma2	1,3

Figure 2. Distribution plot x and y

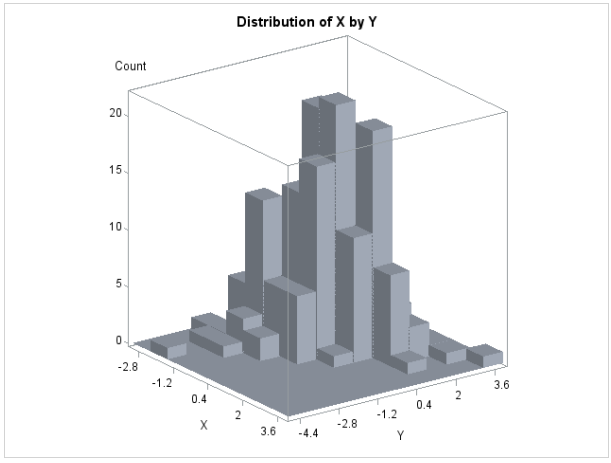
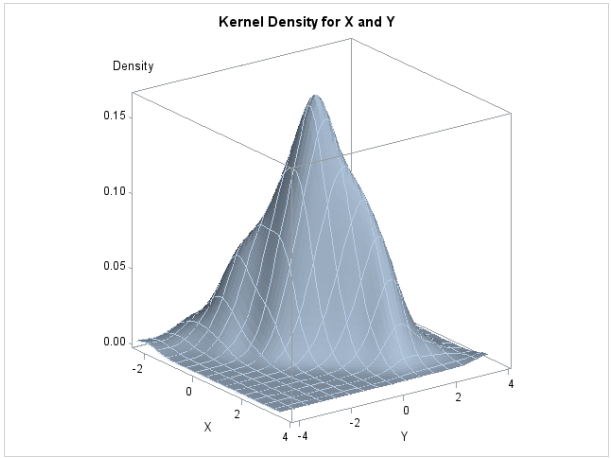


Figure 3. Density plot x and y



From the table 2, showing the simulate parameter $\mu 1$=0.11, $\mu 2$=0.2 and the $\sigma 1$=1.05,$\sigma 2$=1.3. So we get the density distribution.And the figure 2 and figure 3 visualizing its distribution and density.

1.3 changing the bandwidths of kernel density estimation

Suppose we would like a slightly smoother estimate. We can change the value of bandwidth. In the following step, we would change the bandwidth and see what it would happen.

1.3.1 Different bandwidths of kernel density estimation

Figure 4. Contour plot for x and y with different bandwidth

(a) Bandwith of 0.5,0.5 **(b)** Bandwith of 1,1 **(c)** Bandwith of 2,2

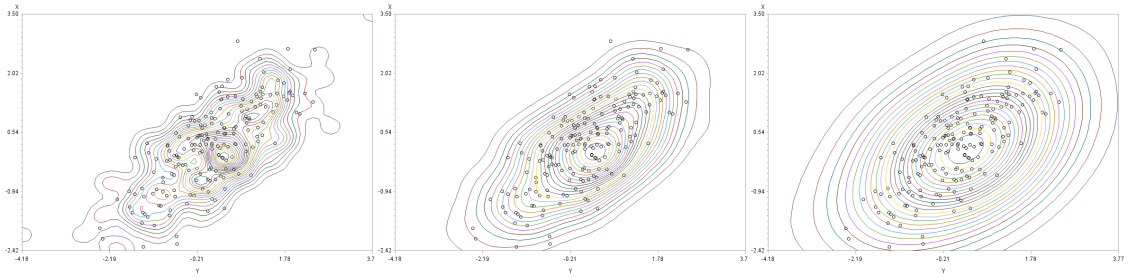


Figure 5. distribution density 3Dplot for x and y with different bandwidth

(a) Bandwith of 0.5,0.5 **(b)** Bandwith of 1,1 **(c)** Bandwith of 2,2

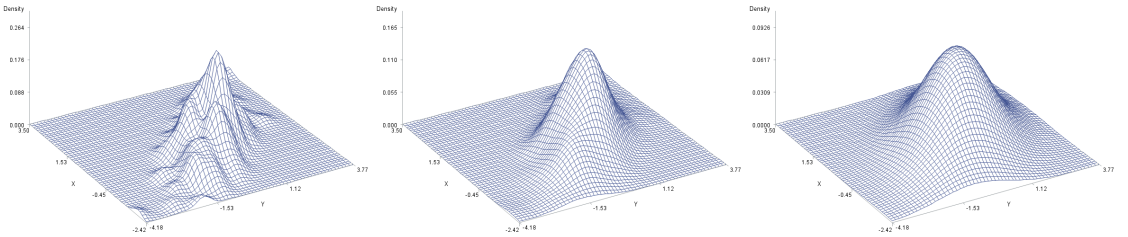


Figure 4 displaying the contour plot with different bandwidth and the figure 5 showing the density 3D plot of distribution. With different value of bandwidth, the contour are very different. With the smaller bandwidth Bandwidth is defined as a range within a band of frequencies or wavelengths. Bandwidth is also defined as the amount of data that can be transmitted in a fixed amount of width. The more smaller bandwidth, the more specify information of variable. We can see, when bandwidth equal to (0.5,0.5), the contour figure are more thin and multi-center which is not a joint normal distribution. And we can see when bandwidth equal to (1,1), the contour figure is an ellipse with just one center. The width of each line is wider than bandwidth of 0.5,0.5 but is more thiner than (2,2) bandwidth.

1.4 Conclusion

From above analysis. In my opinion, I would like to choose the bandwidth of (1,1). Firstly, bandwidth of (0.5,0.5) is too specify for each point which mean the estimator is not general

for all the point. Secondly, compare to the joint density distribution(figure 3), it is just one center which is more similar to the contour plot with bandwidth(1,1). Thirdly, bandwidth of (2,2) is too large to estimate the variable. It contain many entity place. Fourthly, if the variables are follow with the normal distribution, it would be more easy to analysis in some situation. So the bandwidth of (1,1) is more suitable. Noticing that the contour plot with (1,1) bandwidth is a ellipse which is a shape of bivariate normal distribution. So we can guess that the variables follow a bivariate normal distribution. And it is verify by follow qqplot. Both variables x and y are a line.

Figure 6. QQplot for x and y

(a) QQPlot of x **(b)** QQPlot of y

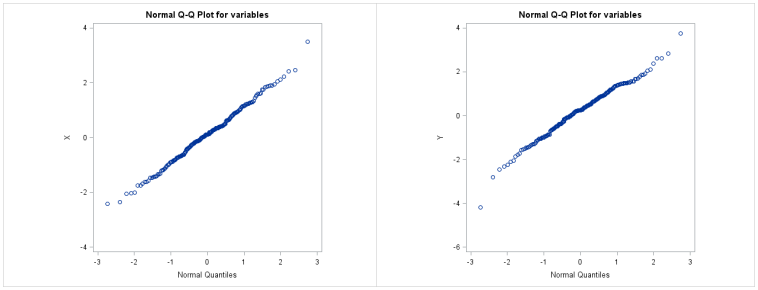
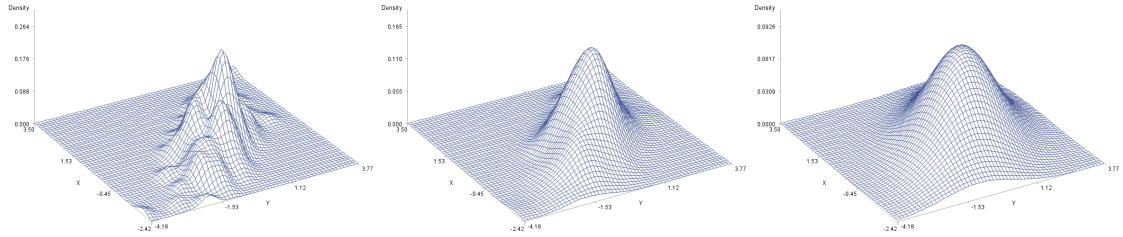


Figure 7. distribution density 3Dplot for x and y with different bandwidth

(a) Bandwith of 0.5,0.5 **(b)** Bandwith of 1,1 **(c)** Bandwith of 2,2



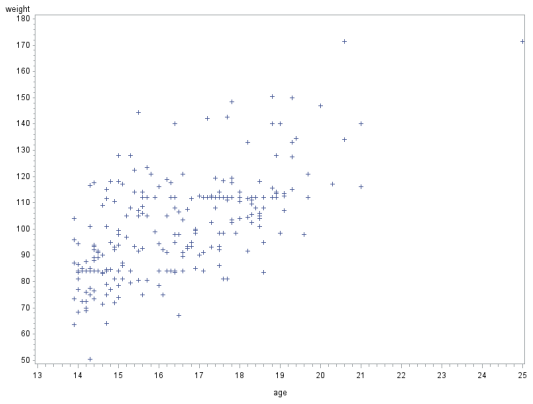
2 Part B

A data set based on the study of 237 children with categorical variable sex and continuous variables age weight and height.We want to exploring the relation between the variables weight and age under different situation. And fit linear regression model between this two variable.

2.1 Pearson's Correlation Coefficient and its p-value

First of all let's consider the relation between the variables weight and age.

Figure 8. plot of two variables



From the above figure we can see the two variable follow a linear relation. To further confirm their relation. In the following step, we will check its Pearson's correlation coefficient.

2.1.1 weight vs. age

Table 3. Pearson's correlation coefficient.

Pearson's correlation	p-value
0,634 64	<0,0001

When we only compared weight with age, its Pearson's correlation coefficient is 0.63464 and its p-value is smaller than 0.0001 which is strongest verify that this two variable have linear relation.

2.1.2 weight vs. age by controlling height

Table 4. Pearson's correlation coefficient.

Pearson's correlation	p-value
0,274 31	<0,0001

When we compared weight with age by set the partial variable height, we get the Pearson's correlation coefficient is 0.23605 and its p-value is smaller than 0.05 which is against the original hypothesis and accept the alternative hypothesis that this two variable have linear relation.

2.1.3 weight vs. age by controlling height with different sex

Table 5. Pearson's correlation coefficient.

sex	Pearson's correlation	p-value
female	0,236 05	0,0130
male	0,314 30	0,0004

When we compared weight with age by set the partial variable height with different sex, we get the Pearson's correlation coefficient of female is 0.23605 and its p-value is smaller than 0.05 which is against the original hypothesis and accept the alternative hypothesis that this two variable have linear relation. we get the Pearson's correlation coefficient of male is 0.31430 and its p-value is smaller than 0.05 which is against the original hypothesis and accept the alternative hypothesis that this two variable have linear relation.

2.2 Simple linear regression between weight and age

Table 6. Parameter Estimates.

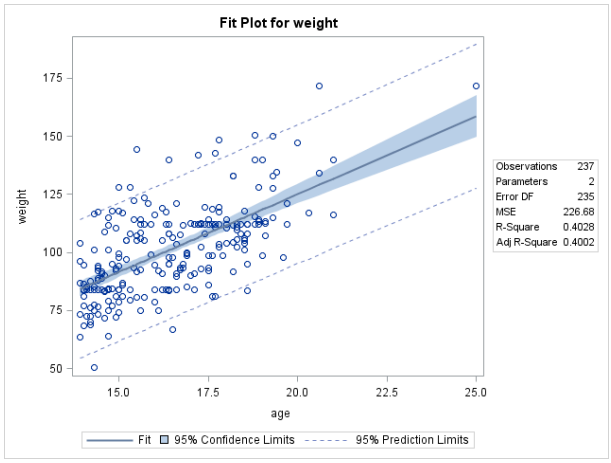
Variable	Parameter Estimate	Standard Error	p-value
intercept	−2,793 51	8,880 047	0,3187
age	6,695 94	0,531 89	<0,0001

Table 7. Analysis of Variance

Source	value
model	p-value<0,0001
R-square	0,4002

From the above table we got the basic summary of fitting model. By using the simple linear regression method, the model is weight = 6.69594 ∗ age -8.79351. The p value of age is smaller than the 0.05 so it is significant. But the p-value of intercept is greater that 0.05 and the r square is just 0.4002, so we have the reason to believe that the simple regression method is not that good fit for this two variable.

Figure 9. plot of two variables



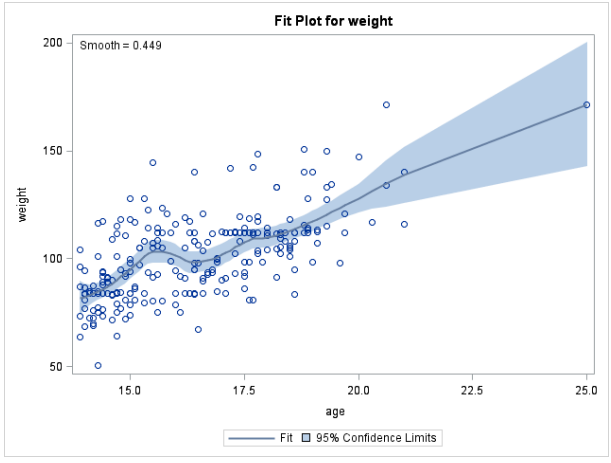
The figure 7 displaying the fitting line and the shape part representing the 95 percent interval. We can see the there are still a lot of scatter plot out of the interval.

2.3 Local regression between weight and age

Table 8. Optimal Smoothing Criterion.

Variable	Value
Smoothing Parameter	0,449
AICC	6,43

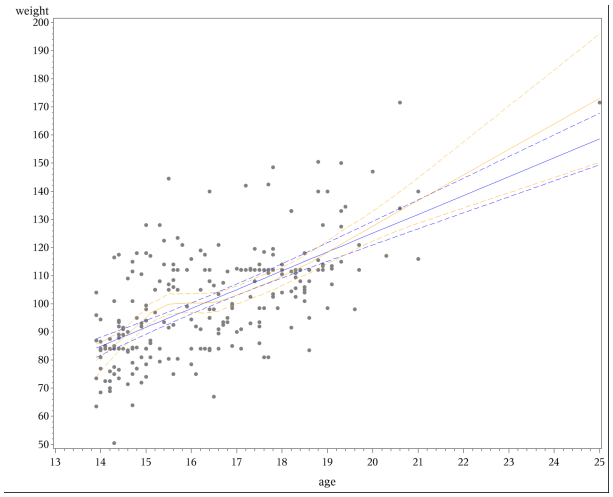
Figure 10. plot of two variables



From the above table and figure, we get the result of local regression. The idea of local regression is that at a predictor x, the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x_i.And we can see the interval of local regression. The fitting line is no longer the straight line but a curve. The local regression is better than simple regression.

2.4 Conclusion

Figure 11. scatter of two fitting model



From the Pearson's correlation coefficient, there are correlation between weight and age. And even under the condition of height and sex, there are still have correlation between them. From the confident interval, the simple regression and local regression are a little bit different. But in general, they are mostly overlay. Based on the interval, we have the reason to believe that they are all follow a linear regression.