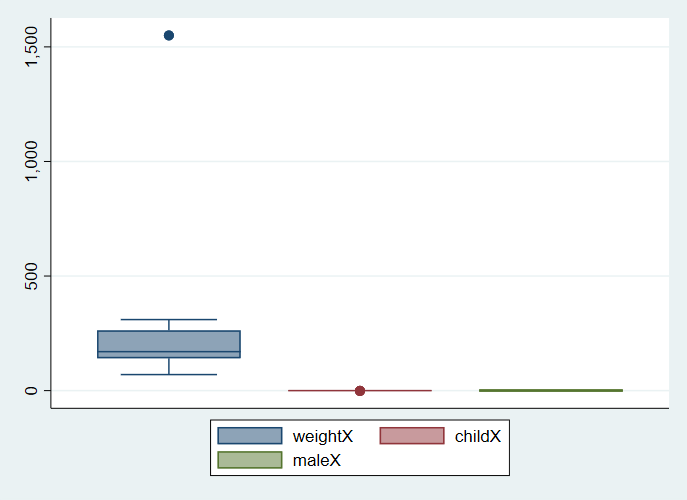
**Take Home Exam#1 Word File**

**For this exam, I make the following truthful statements:  
 1. I have not received any non-instructor approved assistance, I have not given any noninstructor approved assistance to another student taking this exam, including discussing the exam with students in another section of the course.  
 2. I did not plagiarize someone else’s work and turn it in as my own  
 3. I understand that acts of academic dishonesty may be penalized to the full extent allowed by the University at Buffalo Student Conduct Code, including receiving a failing grade for the course with a transcript notation and being expelled from the university. I recognize that I am responsible for understanding the provisions of the University at Buffalo Student Conduct Code as they relate to this academic exercise.**Name: Hao Wu Sign: *Hao Wu* Date: 02/22/2021

Question 1a  
First of all we need get Summary the dataset type and review missing value

This dataset has 4 variable labels 'dounuts','weight','child','male'. Variables donuts' type is str, it should be float. After checking the list of donuts table find out a str value 'five'. We need to replace 'five' to 5, and change variable donuts type to float. Then from the box plot, we regonise the weight of Patty doesn't make sense. Through review variable child, we know it is indicate variable, it should be int and equal to 0 or 1. We need to replace -1 to 1. we create new variable to change value

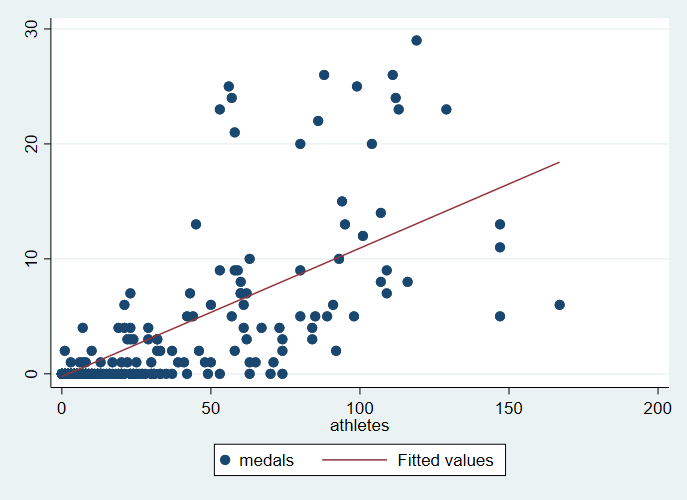
Question 1b

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| donutsx2 | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| weightx2 | .0724512 | .0153915 | 4.71 | 0.001 | .0385748 | .1063276 |
| \_cons | -7.004308 | 2.874728 | -2.44 | 0.033 | -13.33154 | -.6770749 |

Question 2a  
It is panel dataset, because it used different countries over time.

Question 2d  
In 1980s,1984s, 1988s,there have 117 observations; In 1992s, there have 113 obervations; In 1994s, there have 110 observations

Question 2e  
There have 5 country have hosted the Winter Olympics, each of them hose 1 time. United States hosted in 1980, Yugoslavia hosted in 1984, Canada hosted in 1988, France hosted in 1992, Norway hosted in 1994

Question 2f  


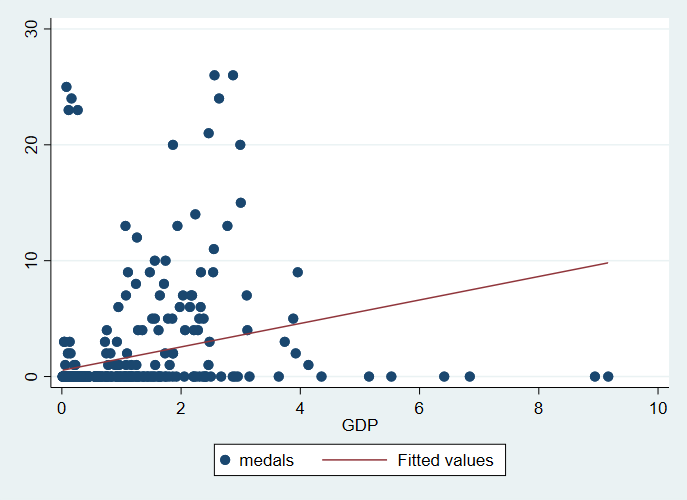
From the graph we could know when the number of athletes is less than 50, the number of medals will not be much. But when the number of athletes is greater than 50, the probability of winning more than 10 medals increases. Generally, the realationship bewtween medals and athletes is positve

Question 2g

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| medals | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| athletes | .1117013 | .004217 | 26.49 | 0.000 | .1034186 | .1199841 |
| \_cons | -.2374701 | .1295405 | -1.83 | 0.067 | -.4919031 | .0169629 |

We could easier get conclusion from table which above this paragraph, the coef is positive and P value is 0, so that there have sufficient evidence that having more athletes cases the medal cout to go up

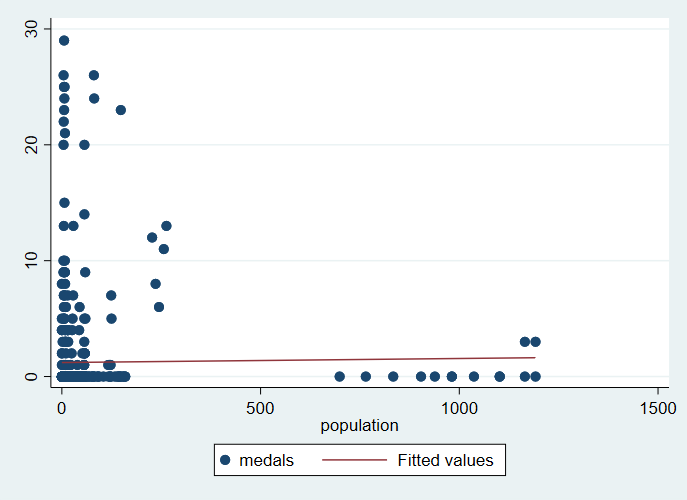
Question 2h



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| medals | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| GDP | 1.01312 | .1609657 | 6.29 | 0.000 | .6967766 | 1.329464 |
| \_cons | .5309645 | .2224805 | 2.39 | 0.017 | .0937268 | .9682021 |

The hypothesis of H0 is: all predictors will not affect y, that is, the coef of all predictors will be 0, and all predictors will not be significant. The value of Prob>F is the probability of the establishment of the above-mentioned H0 hypothesis. When it approaches 0, it means that at least some predictors’ coef is not 0.

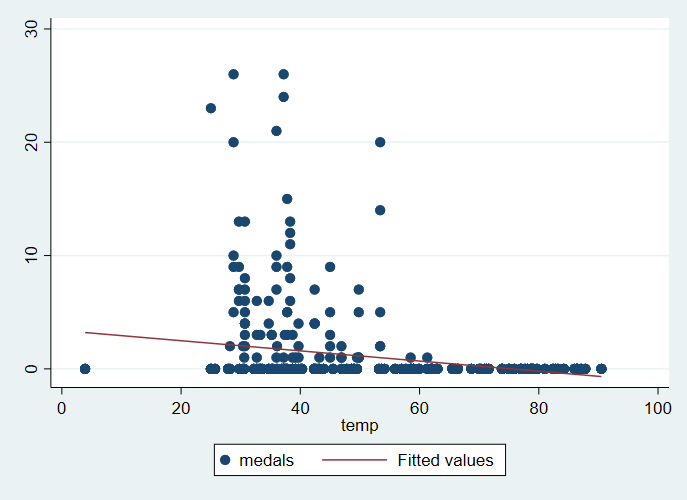
Question 2i



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| medals | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| population | .000349 | .0010801 | 0.32 | 0.747 | -.0017725 | .0024705 |
| \_cons | 1.208663 | .1779074 | 6.79 | 0.000 | .8592262 | 1.5581 |

only when p value approch to 0, we have effident to reject H0, after we run the regression, we know p value is 0.7467, significant larger than 0, so that we cant reject the slope coefficient.

Question 2j



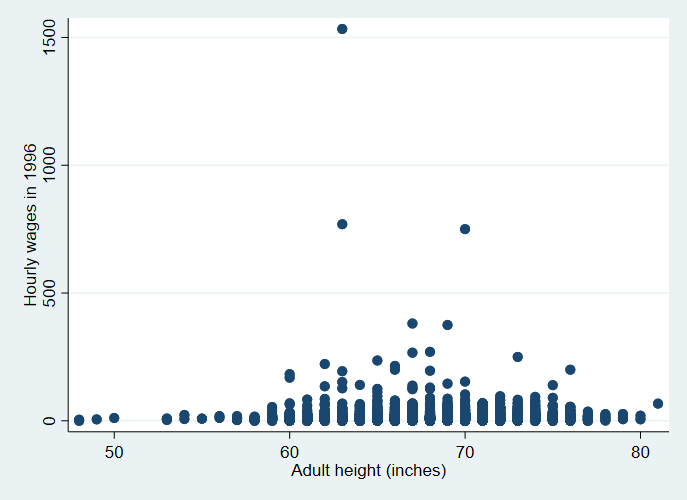
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| medals | Coef. | Std. Err. | t | P>|t| | [93% Conf. Interval] | |
| temp | -.0448618 | .006971 | -6.44 | 0.000 | -.0575171 | -.0322064 |
| \_cons | 3.383477 | .4000357 | 8.46 | 0.000 | 2.657242 | 4.109712 |

The conef is - 0.0448618. The P-value is 0 less than 0.07, so that we should reject H0. The variance have statistic significant. The confident interval is -0.0575173 to -0.0322064

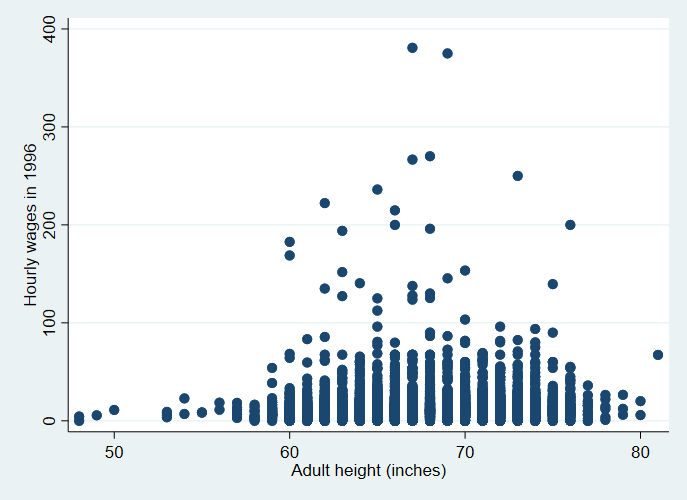
Question 3a

When we got large sample dataset, we could use mean of each variable to replace the missing value

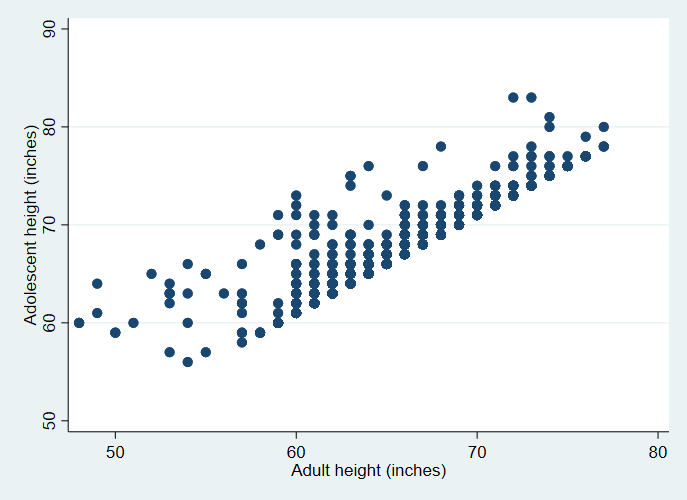
Question 3b

  
We can see that there are three noise point belonging to the outline value, but because our sample is large, it will not affect our next behavior.

Question 3c

  
These data show a normal distribution trend, and most of the values are concentrated in 60-75

Question 3d

  
Is not statistically significant, so we don’t use

Question 3e  
We need to calculate the average of the samples by gender, and then classify whether they belong to the label of tall according to gender. All missing values are represented by

Question 3f

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| wage96 | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| tall | 1.301808 | 1.255708 | 1.04 | 0.300 | -1.160269 | 3.763884 |
| \_cons | 12.86399 | .84967 | 15.14 | 0.000 | 11.19803 | 14.52994 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| wage96 | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| tall | 1.092977 | .6136992 | 1.78 | 0.075 | -.1102646 | 2.29622 |
| \_cons | 14.39368 | .4210689 | 34.18 | 0.000 | 13.56811 | 15.21924 |

P values are all greater than 0.05, so there is no statistical significance, not reject H0.

Question 3g

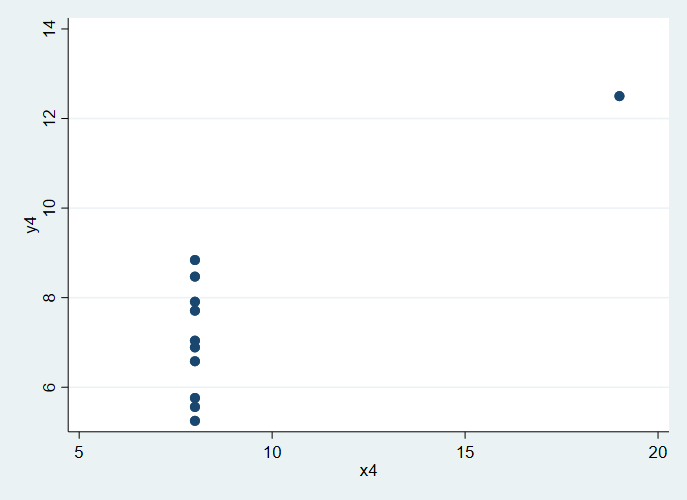
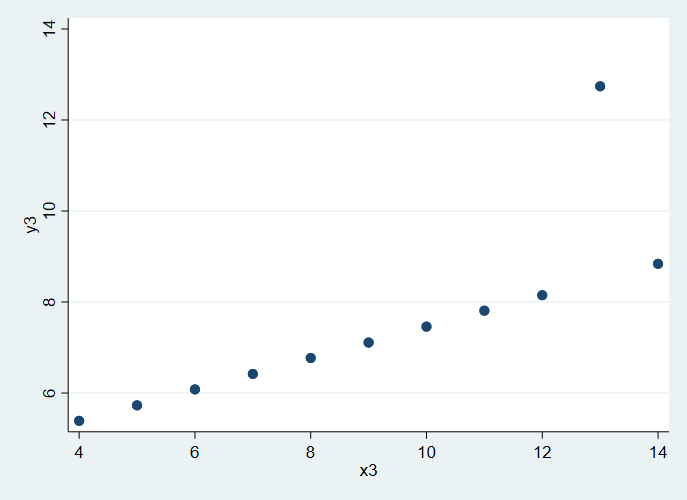
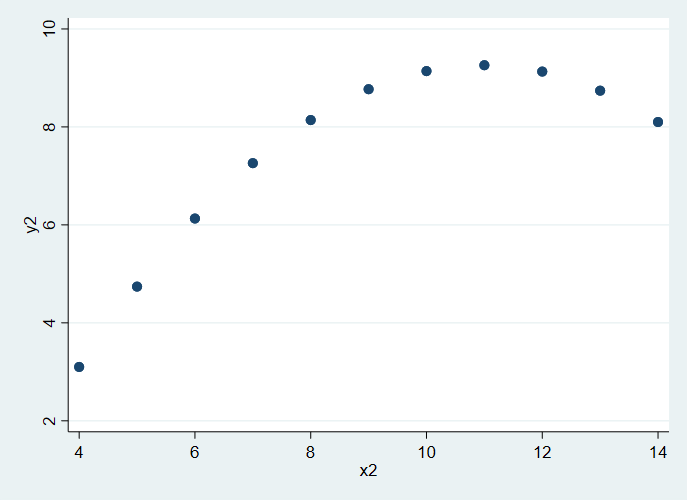
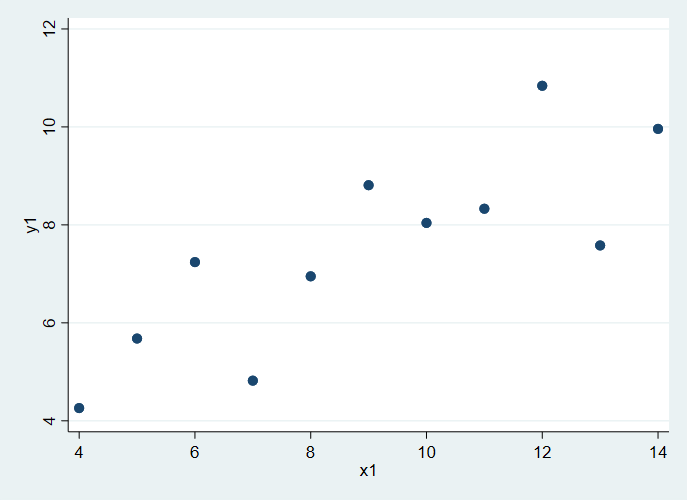
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| wage96 | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| tall | 2.619421 | .6590893 | 3.97 | 0.000 | 1.327139 | 3.911703 |
| \_cons | 11.54637 | .4461102 | 25.88 | 0.000 | 10.67168 | 12.42106 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| wage96 | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] | |
| tall | 1.487828 | .4488337 | 3.31 | 0.001 | .6078274 | 2.367828 |
| \_cons | 13.99883 | .3079957 | 45.45 | 0.000 | 13.39496 | 14.6027 |

All P values tend to be close to 0, so there is significant statistical significance, reject H0.

Question 4a

All X have same mean and variance, all Y also have same mean and variance.

Question 4b  


Question 4c  
The first is a positive correlation, the second is a normal distribution, the third is a linear positive correlation, and the fourth two variables are independent