**HW 09**

**Fangling Zhang**

**Q24**

The bias component for the  fitted value is .

Here  is the true mean respoense when the levels of the predictor variables are those for the icase, and in this case we use to calculate.

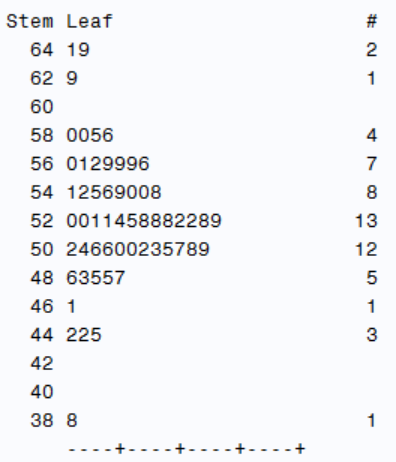
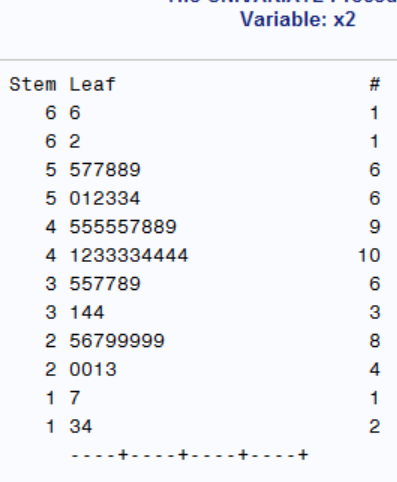
is the expection of the  fitted value for the given regression model, and in this case we use to calculate.

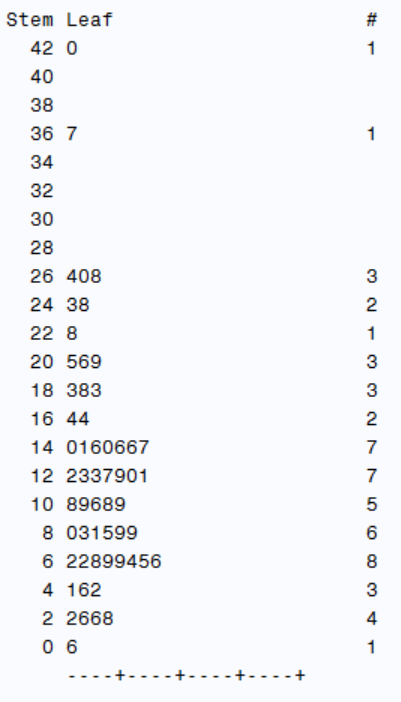
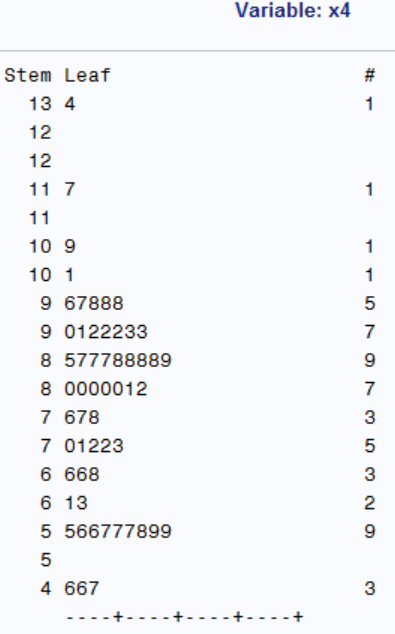
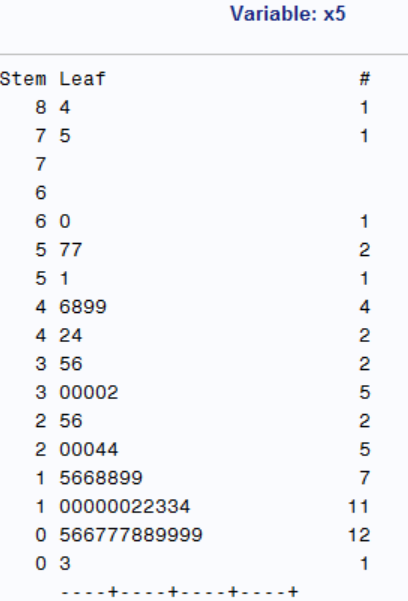
The sampling error component for is: . In this case, we use  to calculate.

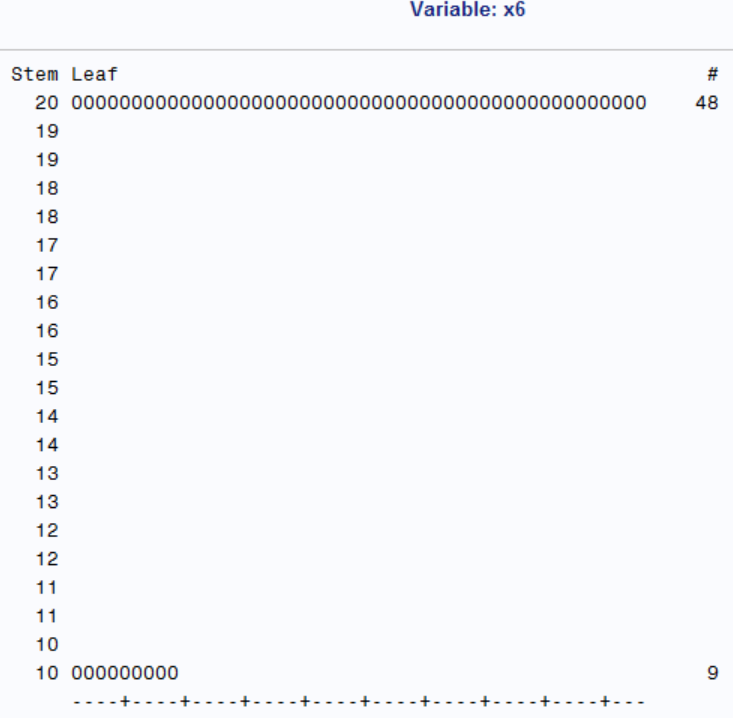
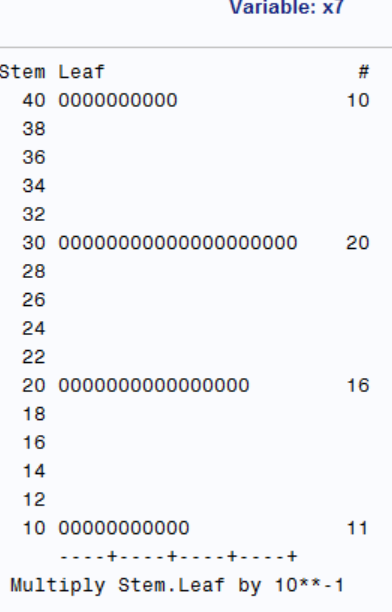
For, . ==10+45\*10=460. =13+40\*10=413. The bias component here =460-515=-55, and the sampling error component here =413-460=-47.

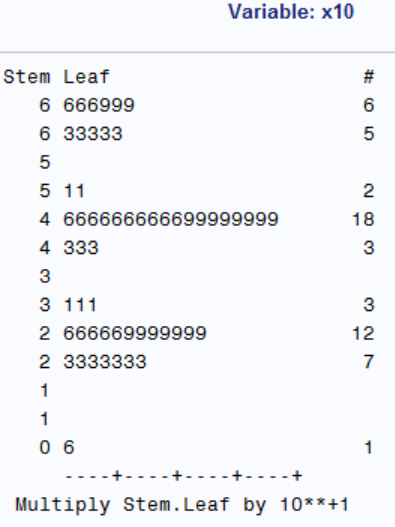
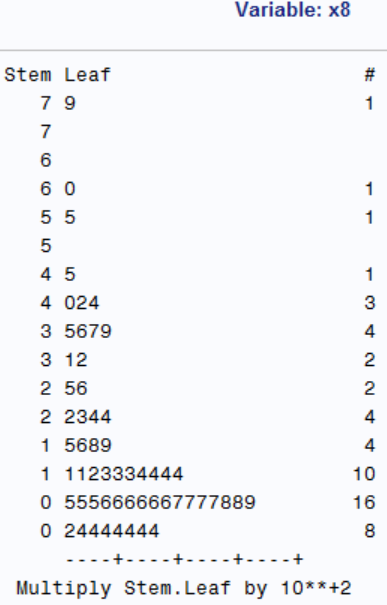
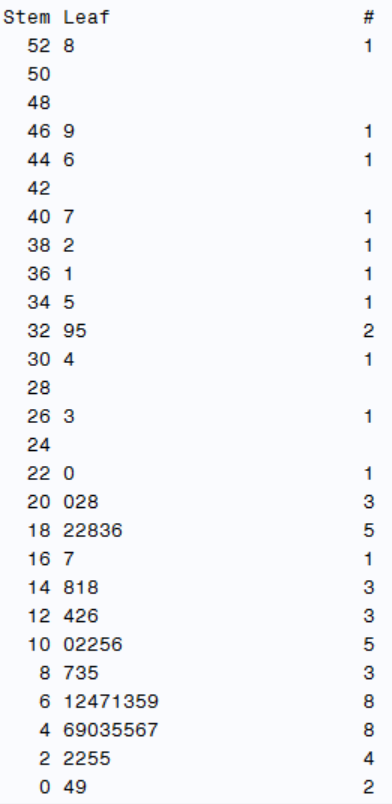
For , . ==10+45\*20=910. =13+40\*20=813. The bias component here =910-1615=-705, and the sampling error component here =813-910=-97.

**Q25**

(a) Stem and leaf plots for each of the predictor variables are as follows:







From the plots above, we can see that x6(Medical school affiliation) just has 2 values, x7(Region) has 4 values. The distributions of x1(Age), x2(Infection risk), x5(Number of beds), x8(Average daily census) and x10(Available facilities and services) are more centralized than those of x3 x4 and x9.

**(b)**

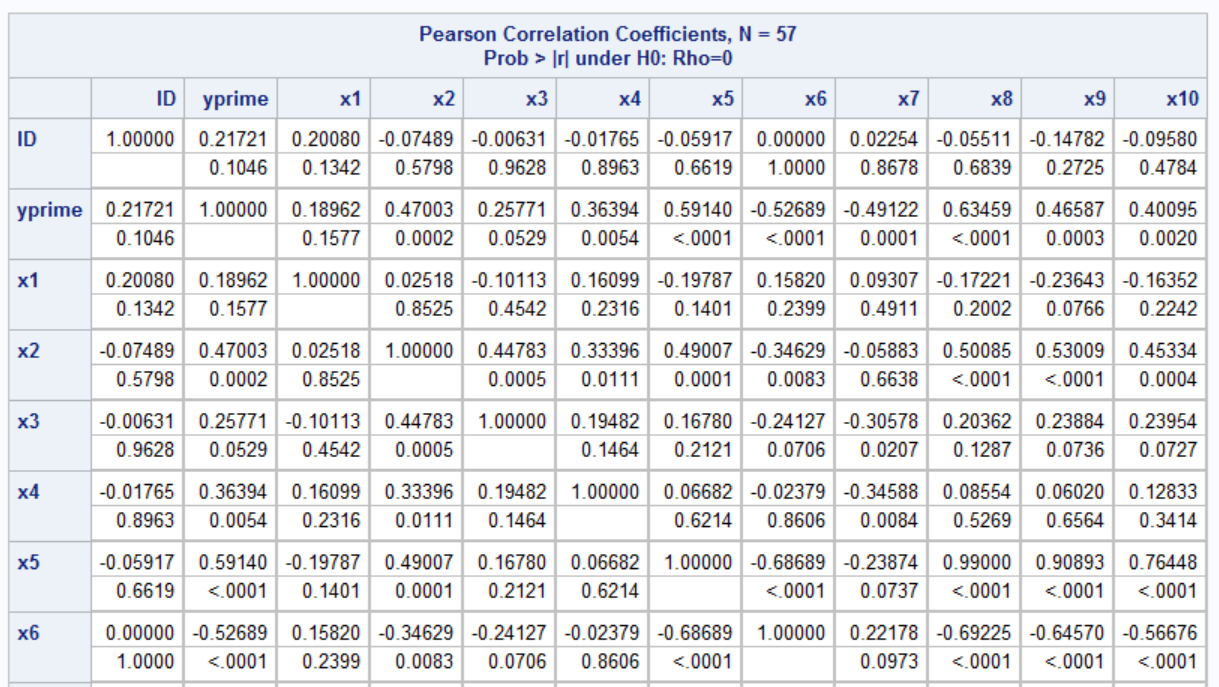
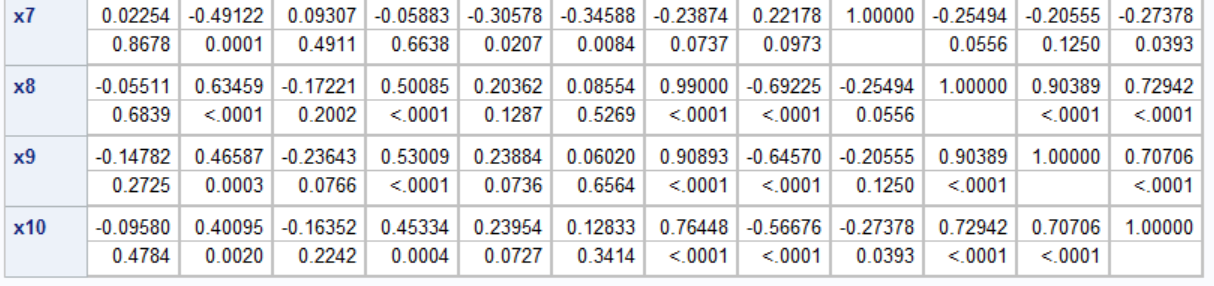
The scatter plot matrix and the correction matrix of the X variables are as follows:

We can see that yprim(Length of stay) are obviously correlated with x5(Number of beds), x6(Medical school affiliation) and x8(Average daily census).

X6(Medical school affiliation), x8(Average daily census), x9(Number of nurses) and x10(Available facilities and services) are obviously correlated with each other.

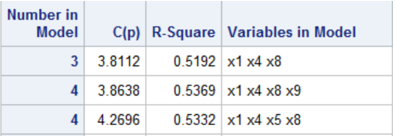
x2(Infection risk), is obviously correlated with x8(Average daily census), x9(Number of nurses).





**(c)**

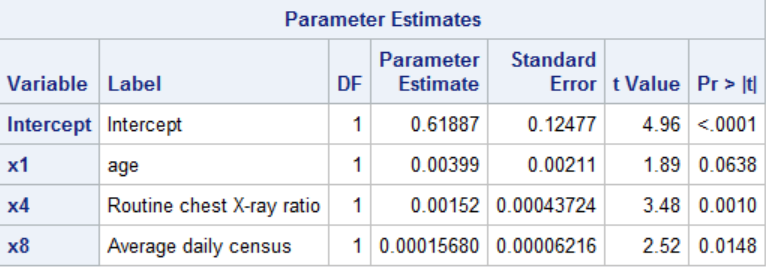
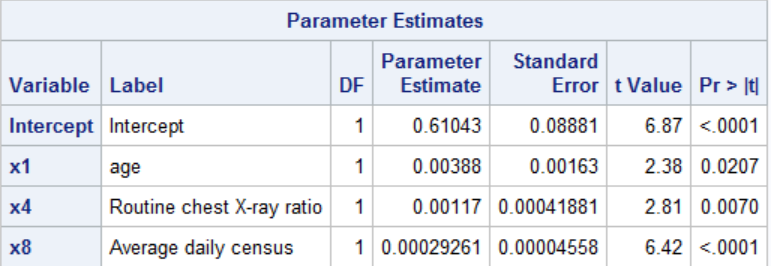
The tree best subsets models according to the Cp criterion are as follows. In these subsets models, the one with the smallest bias is the first one with variables x1 x4 x8.

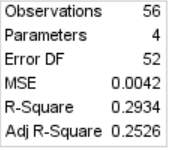
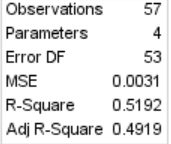


**Q27**

(a)

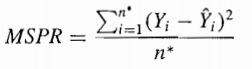
The left table is the regression model fit on the validation data set, and the right one is the best model identified in Project 9.25. Obviously the estimated regression coefficients, standard deviations, the error mean squares and coefficients of multiple determination are all different between these two models.





(b)

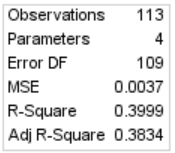
The mean squared prediction error

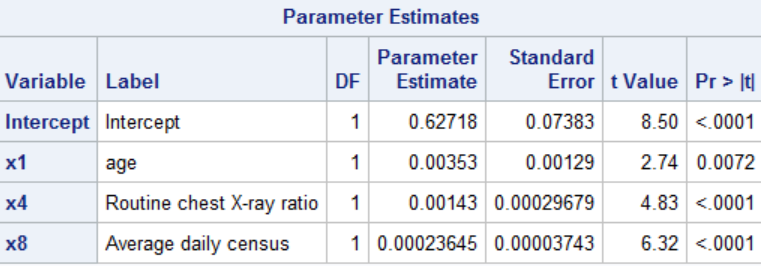




We get MSPR=0.0046, the MSE obtained from the model-building data set is 0.0031. As mspr and MSE are close, so there seems to be no evidence of a substantial bias problem in MSE here.

(c)

Combine the two data sets and fit the selected regression model to the combined data. The results are shown as follows:



We can see that the estimated regression coefficients and their estimated standard deviation different from those for the model-building data set. The differences in the estimates are expected. Because the estimates of models on the model-building datasets and validation datasets are different, when the validation datasets are appended, the estimates of original model change.