Multiple Regression Revenue Homework

 For this assignment, I thought it would be best to use the pairs command to get a look at all the correlations and see which were the strongest with SalesAmt on their own to get an idea of which to choose for my Multiple Regression model. I used the command ‘pairs(Revenue, pch=’\*’’ to produce the following correlogram

Then, after reading further into the assignment and watching the instructional video over a few more times, I decided that this was not the best course of action to take. Instead, if my goal is to find the model that is the best at predicting, I felt that I should use the significance levels provided by the ‘summary’ command to determine which x values needed to be in my model. I used the command ‘allxmodel=lm(SalesAmt~Employee.turnover+No.Employees+No.SalesCampaigns+RDExpense+No.NewProdToMarket+MaterialCost+No.CustComplaints+No.Cust, data=Revenue)’ to produce a model that included every single possible x value. The summary command (summary(allxmodel)) informed me that the following were the only values with any level of significance in predicting SalesAmt: No.CustComplaints (p<.0001), Employee.turnover (p<.05) and No.Employees (p<.05). Then, I moved forward with the analysis by trying different combinations of these variables as nonlinear relationships. When tested as a nonlinear relationship, Employee.turnover had a higher p value than when it was tested as a linear relationship. However, No.Employees, when tested as a nonlinear relationship has a p value that was now less than .0001! Interestingly, No.CustComplaints still had a highly significant p value when tested as a nonlinear relationship; so, I thought it would be best to hone in on the rest of my variables, and then try my ‘final’ model with No.CustComplaints as both a linear and nonlinear variable, and then compare the two via an F-Test or an AIC.

Now that I had figured out which of the ‘big three’ were more significant predictors of SalesAmt as either linear or nonlinear variables, I *almost* thought it was time to test my theory on No.CustComplaints as either a linear or nonlinear relationship. Then I thought, ‘what if some of the variables I omitted earlier on were actually significant if they were tested in a nonlinear relationship?’ Thus, I used the following command: ‘waitaminute= lm(SalesAmt~I(No.SalesCampaigns^2)+I(RDExpense^2)+I(No.NewProdToMarket^2)+I(MaterialCost^2)+I(No.Cust^2), data=Revenue)’ to determine if any of the other variables would have significant p values if the relationship between them and SalesAmt was tested as nonlinear. Using this command, I found that No.SalesCampaigns had a p value of <.0001 when tested as a nonlinear relationship! Additionally, RDExpense now had a p value of <.01 and No.Cust did as well. Hence, I decided to include all of the above in my (close to) final model.

I then used the command ‘almostthere= lm(SalesAmt~I(No.CustComplaints^2)+I(No.Employees^2)+I(No.SalesCampaigns^2)+I(RDExpense^2)+I(No.Cust^2)+Employee.turnover, data=Revenue)’ to build a model with all of the variables I had found to have a significant relationship with SalesAmt thus far. However, I soon found out that some of the variables no longer had a significant p value when they were all combined in model with so many x variables. At this point, I decided to go back and build a model with only variables that had a p value of p<.0001.

Of course, I had tested so many models previously, I didn’t actually need to recreate any more models; I could just go through the console file and see which models had the highest R2, the most variables with significant p values and the lowest standard error, and compare them using either the ‘anova’ command or the ‘AIC’ command. Thus, I decided to compare the following models using AIC:

lm(SalesAmt~ I(No.Employees^2)+I(No.CustComplaints^2)+I(No.SalesCampaigns^2), data=Revenue)

lm(SalesAmt~I(No.CustComplaints^2)+I(No.Employees^2)+Employee.turnover,data=Revenue)

lm(SalesAmt~I(No.CustComplaints^2)+I(Employee.turnover^2)+I(No.Employees^2), data=Revenue)

All three had very similar AIC values (631.2546, 631.9486 and 633.2662 respectively). Hence, the first model should be the winner, right? Perhaps, but I still want to check and see if I get different results if I change the No.CustComplaints from a nonlinear relationship to a linear one. So, I ran the AIC command three more times: each time I compared each of the models to a new model, each with ‘I(No.CustComplaints^2)’ changed to ‘No.CustComplaints’. This little experiment yielded that both of the latter models had lower AIC values when No.CustComplaints was changed from nonlinear to linear. In my eyes, this meant that one last AIC command must be ran to see if these new models could stand up to the original champ ‘lm(SalesAmt~ I(No.Employees^2)+I(No.CustComplaints^2)+I(No.SalesCampaigns^2), data=Revenue)’.

Wait for it…

The revamped models beat the original champ! The second model [lm(SalesAmt~I(No.CustComplaints^2)+I(No.Employees^2)+Employee.turnover,data=Revenue)] had an even lower AIC score than the original winner [lm(SalesAmt~ I(No.Employees^2)+I(No.CustComplaints^2)+I(No.SalesCampaigns^2), data=Revenue)] when they were compared head-to-head.

Therefore, I conclude that the best way to predict sales is to use the model ‘[lm(SalesAmt~I(No.CustComplaints^2)+I(No.Employees^2)+Employee.turnover,data=Revenue)’ because it has an adjusted R2 value of 0.9947 which means that 99.47% of the change in sales can be predicted by this model. Also, two of the three variables used have p values that are less than .0001 (the other is less than .05) which means you can be between 95 and 99.999% confident that these findings can be extrapolated to the population in question.