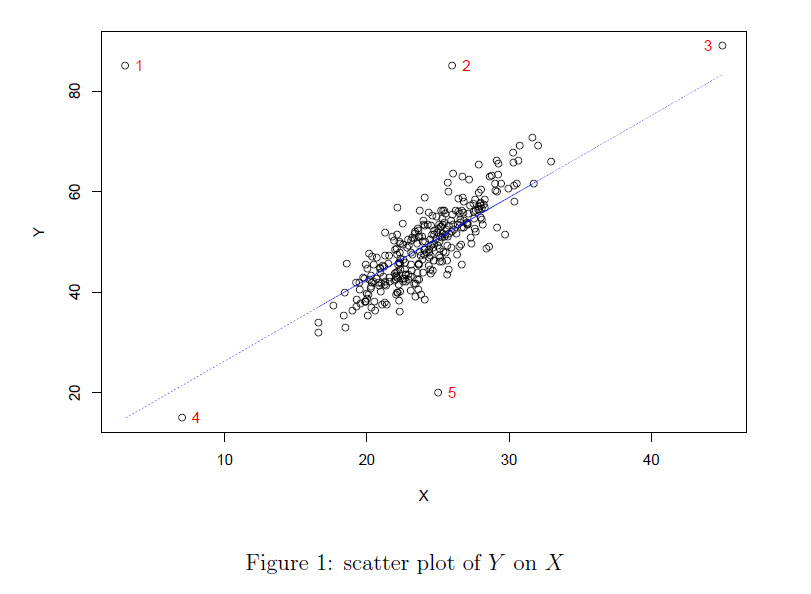
1. To find unusual points in a data set used for simple linear regression Yi = β0 + β1Xi + εi (i = 1, 2, …, 300), a data analyst examines the scatter plot of Y on X as shown in Figure 1. The estimated least square regression line is drawn as well. Sample means X-bar = 24.2, Y-bar = 49.4. Classify each of the 5 unusual points on the plot according to type (i.e., Outlier? High leverage point? Influential point?) and justify your answer.



All points 1-5 are can be classified as outliers, as they have unexpected response value y compared to their expected value, on or near the line of best fit. They points also have the following characteristics:

Point 1 - Outlier and Leverage point. This response value is very different from the expected value as it has a great distance from the line of best fit.

Point 2 - Outlier and Leverage point. This response value is very different from the expected value as it has a great distance from the line of best fit.

Point 3 - Outlier and Influential point.

Point 4 - Outlier only.

Point 5 - Outlier and Leverage point. This response value is very different from the expected value as it has a great distance from the line of best fit.

1. Name one or more graphs that can be used to validate each of the following assumptions.
   1. The error terms have constant variance.

Residual Scatter Plot. This plot is helpful for determining non-constant variance, so by definition, can also assist with identifying constant variance. Error terms with constant variance would not show a specific shape or pattern on the plot.

* 1. The error terms are normally distributed.

Normal probability plot, or Normal Q-Q plot. It will separate the data into basically equal-sized subsets. This graphical tool allows us to assess whether data is normally distributed or not.

* 1. There is a linear relationship between the response and predictor variables.

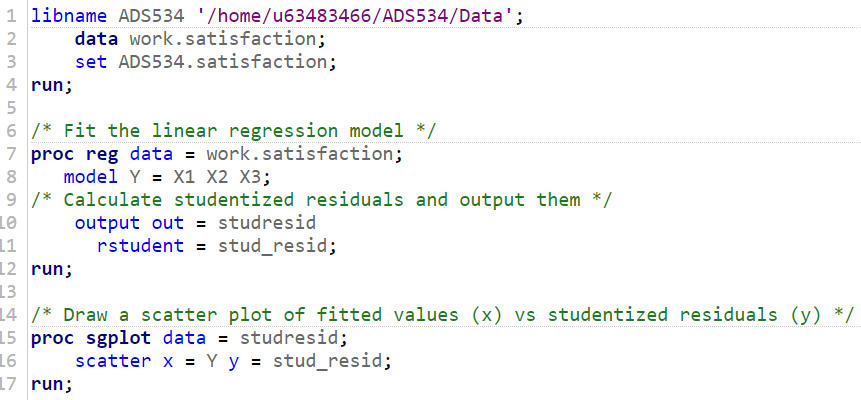
A residual scatter plot. Points can be plotted between the response variable (y) and the predictor variables (x) one at a time, to examine potential linear relationships that would be evaluated visually.

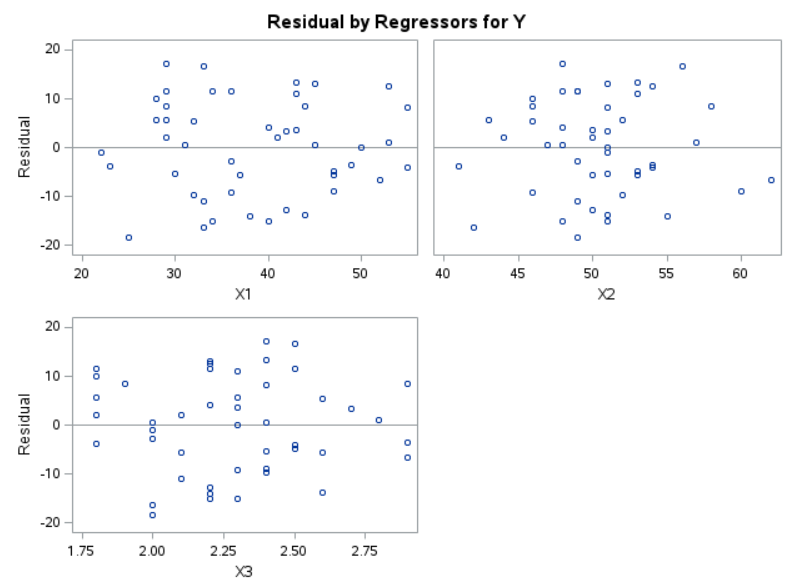
1. KNN, problem 10.11, edited. (page 415) A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X1, in years), severity of illness (X2, an index), and anxiety level (X3, an index). The administrator randomly selected 46 patients and collected the data in satisfaction.sas7bdat, where larger values of Y, X2, and X3 are, respectively, associated with more satisfaction, increased severity of illness and more anxiety.

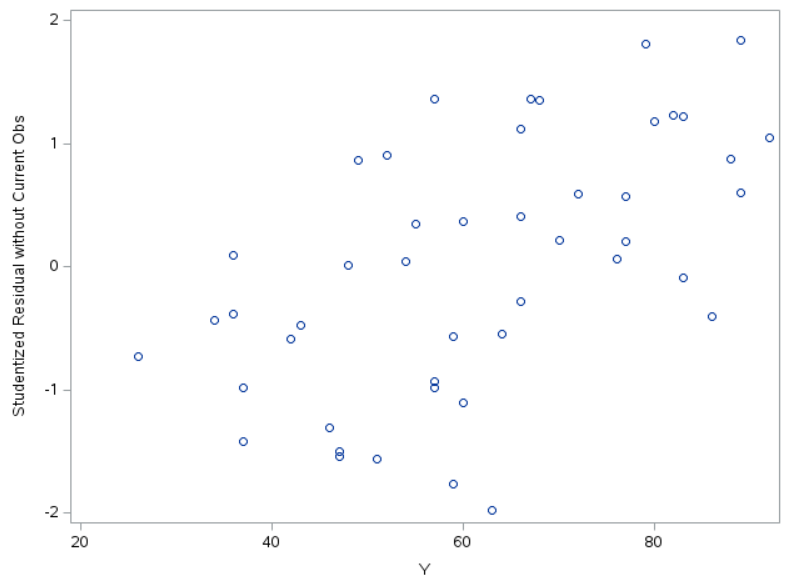
Fit the regression model

Yi = β0 + β1X1i + β2X2i + β3X3i + εi, i = 1, 2, …, 46:

* 1. Using SAS, obtain the studentized residuals and draw a scatter plot where x-axis is fitted value, and y-axis is the studentized residual. Do there appear to be any outliers?

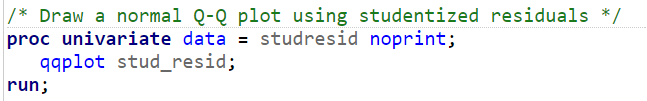


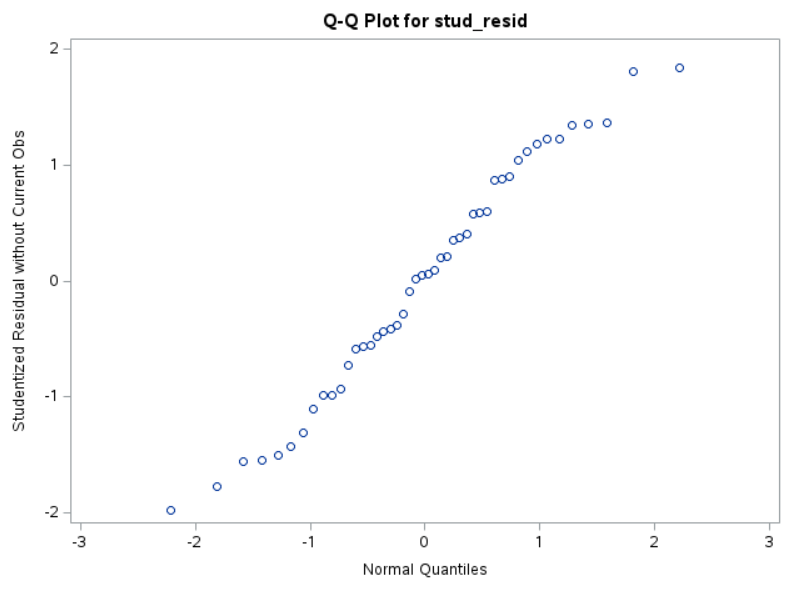




Based on a visual look at the scatter plot, there do appear to be some outliers. Assuming there should be a standard increasing slope of the line of best fit, then some of the points between X = 40 and X = 50 fall lower than would be expected. This would need to be confirmed with further analysis for definite conclusions. Ideally, most of the data would be clustered around the value Y = 0.

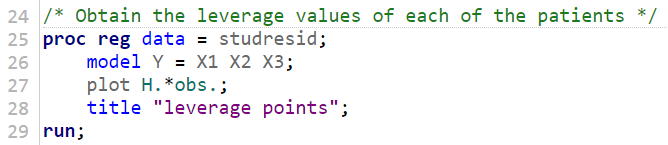
* 1. Draw a normal probability plot (normal Q-Q plot) using studentized residuals. Interpret your plots and summarize your findings.

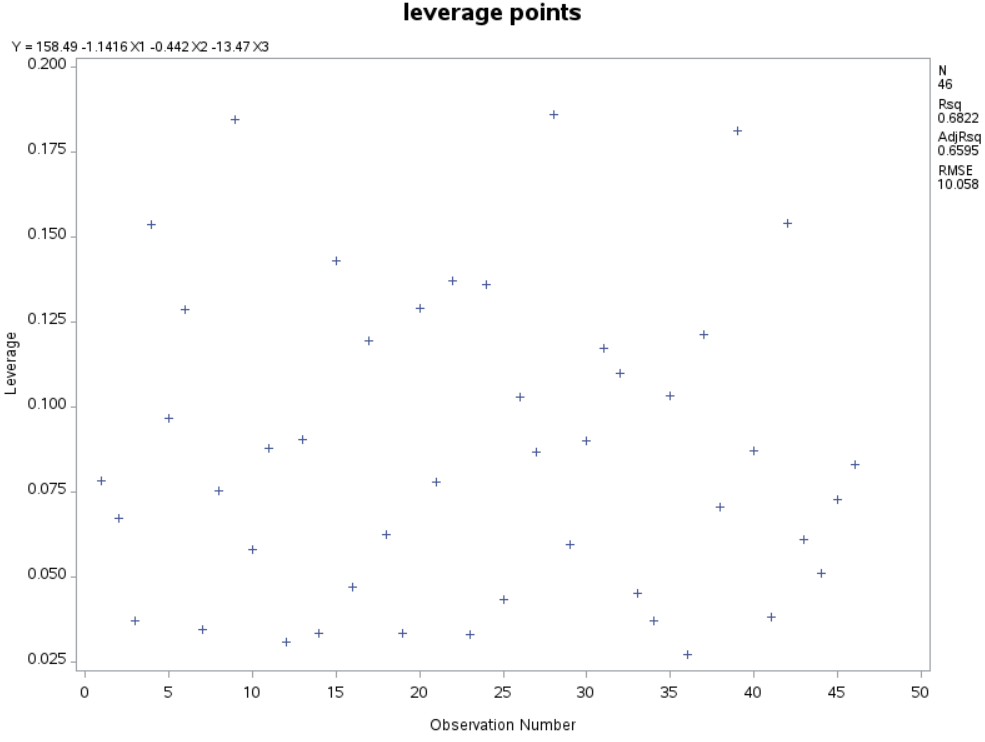




The points mostly lie along a straight line, but there are some deviations, especially at the beginning of the line (around X = -2) and the end of the line (around X = 2). This aligns with what can be expected with normally distributed data.

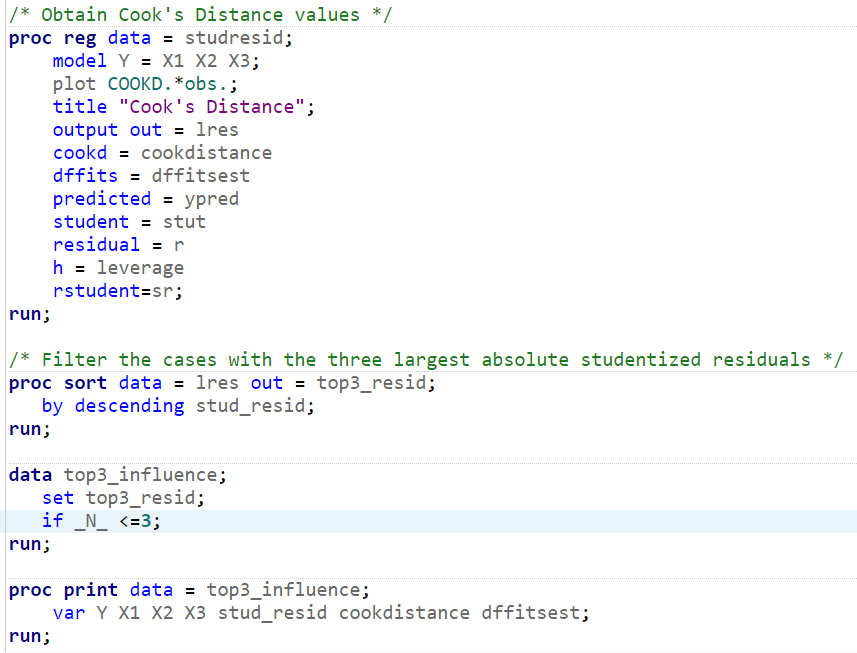
* 1. Obtain the leverage values of each of the patients using SAS. Is there any high leverage point? (using the rule of thumb discussed in class to make conclusion)

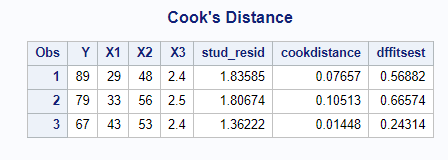




Leverage points outside of ±2(predictors/observations) would fall outside of -0.13 and 0.13. Since we are looking at high leverage points, values above 0.13 are the ones of interest on this plot. It appears that roughly 8 points fall above 0.13 on this scatter plot out of 46 total observations, or roughly 17.4% of the observations.

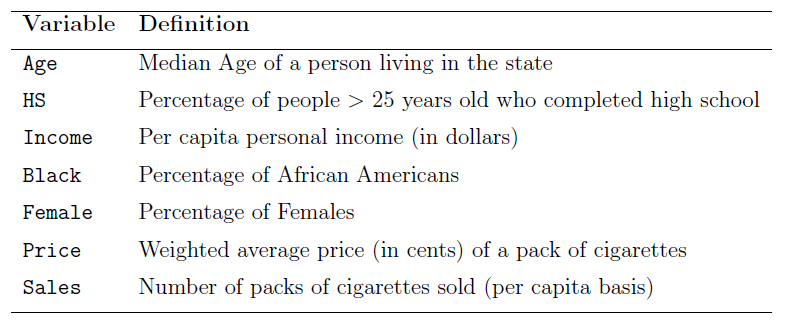
* 1. The three largest absolute studentized residuals are for cases 11, 17, and 27. Obtain the Cook's distance values and DFFITS for these cases to assess their influence. What do you conclude?





The three observations have very small Cook’s distance values, of 0.07657, 0.10513, and 0.01448. Using a cutoff of 4/n, where n is the total 46 observations from the original dataset, we could use a cutoff of 0.087. Under that cutoff, we could conclude that at least two, if not all three, observations are not influential on the overall model. For the dffitsest, if we use 2√(p/n), our cutoff for acceptance would be 0.5108. In this case, at least one value is small enough to not be considered influential on the predicted values when removed.

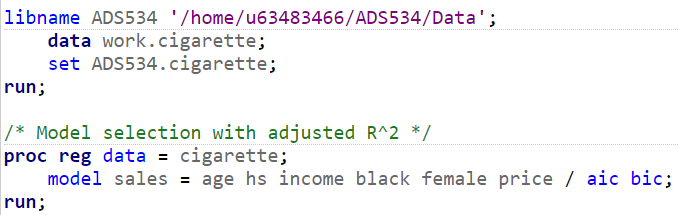
1. A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The variables for the study are as follows:

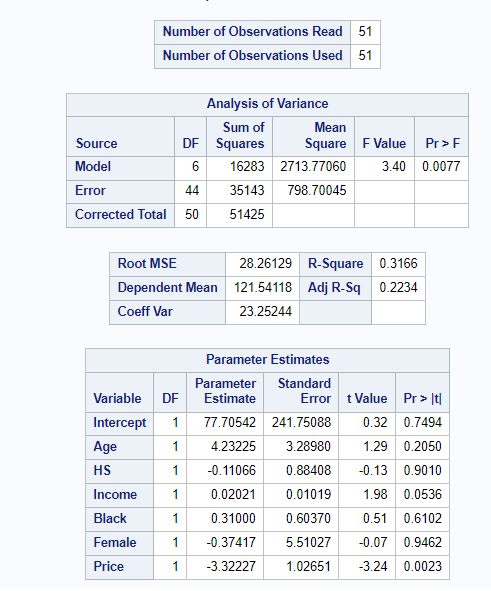


The full data set is available in Moodle (cigarette.sas7bdat).

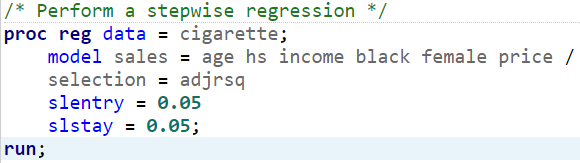
The investigators are interested in studying how statewide cigarette consumption is associated with various socioeconomic and demographic variables, and building a parsimonious regression model for predicting the consumption of cigarettes.

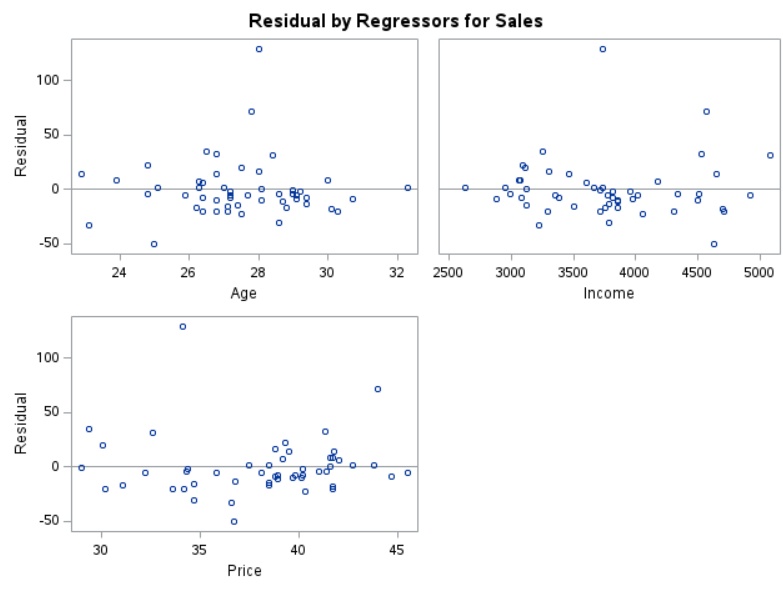
* 1. Build a best linear regression model that predicts the per capita sale of cigarettes in a given state. Perform your analysis using variable selection criterion adjusted R2. Which variables are included in the final model?





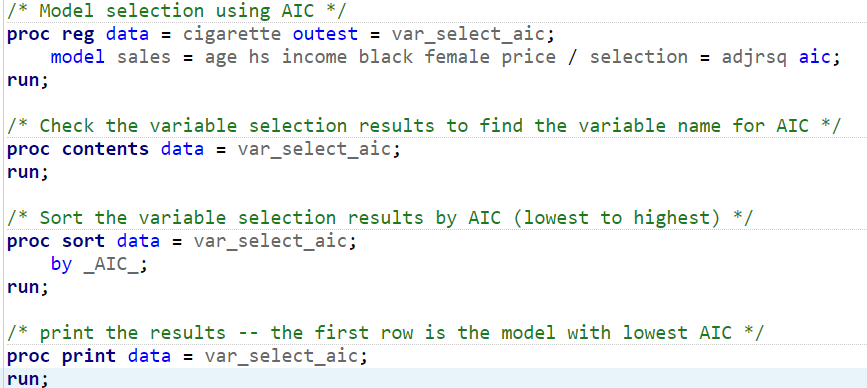
Based on the output of the model selection, we will use stepwise selection.





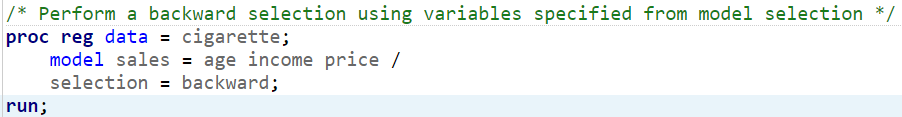
After specifying entry and stay criterion of p < 0.05 and the adjusted R2 value, the Age, Income, and Price variables make it into the final model.

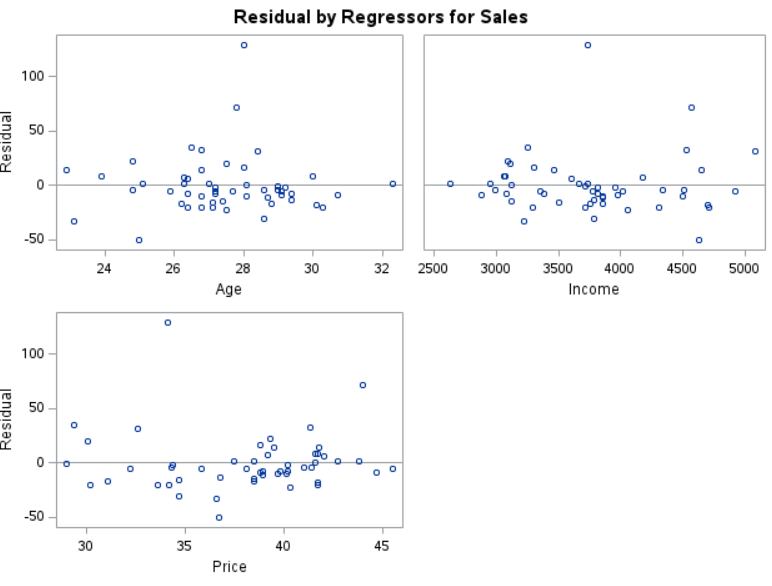
* 1. Build a best linear regression model that predicts the per capita sale of cigarettes in a given state. Perform your analysis using variable selection criterion AIC. Which variables are included in the final model?





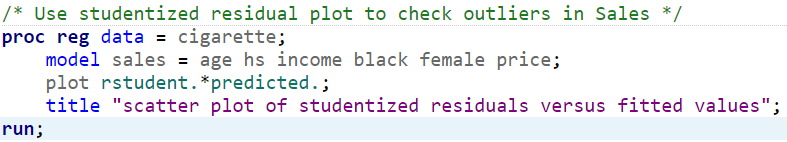
Based on the output of the model selection, we will use backward selection.

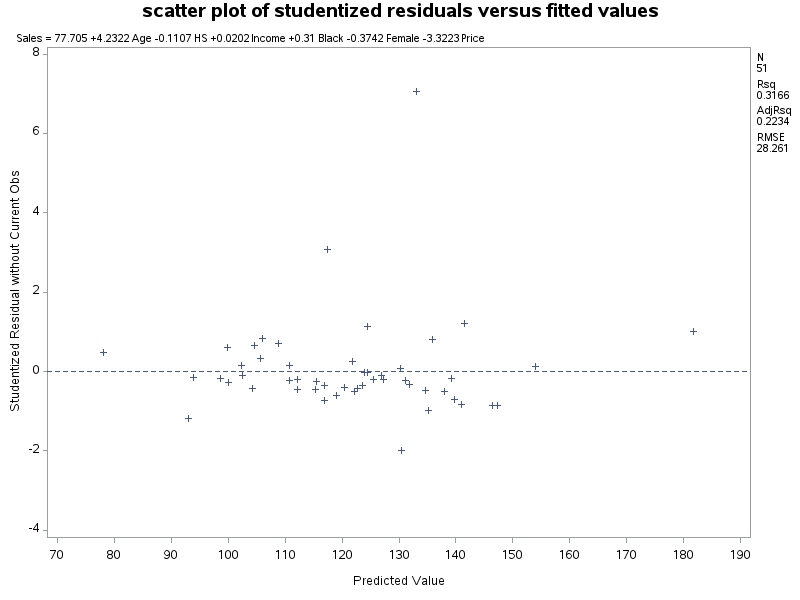




Based on the model selection by AIC, the variables are specified by the output. Age, Income, and Price make it into the final model. Then a backward selection is used to generate the final model.

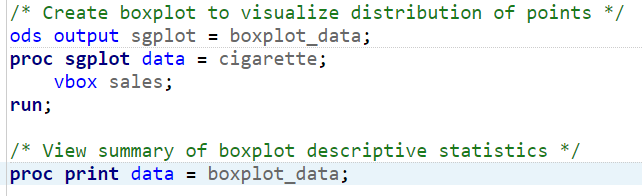
* 1. Use studentized residual plot to check whether there is any outlier in response variable Sales?

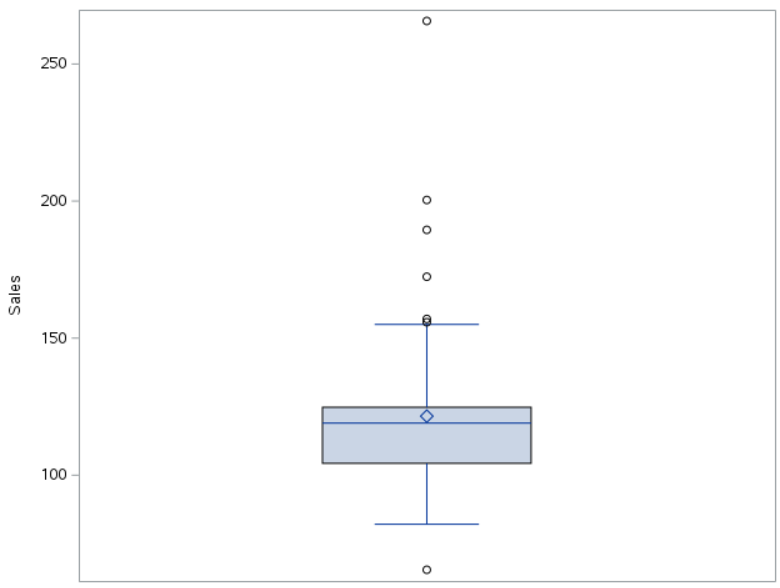


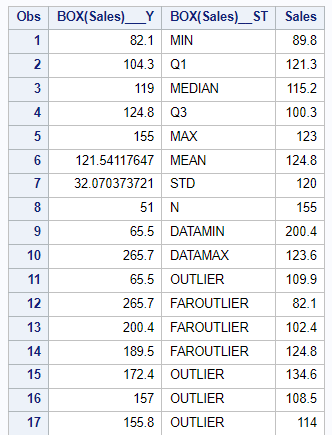


There do appear to be outliers in the response variable Sales.

* 1. Delete any outlier you find. Redo part (a). Do you obtain the same final model?







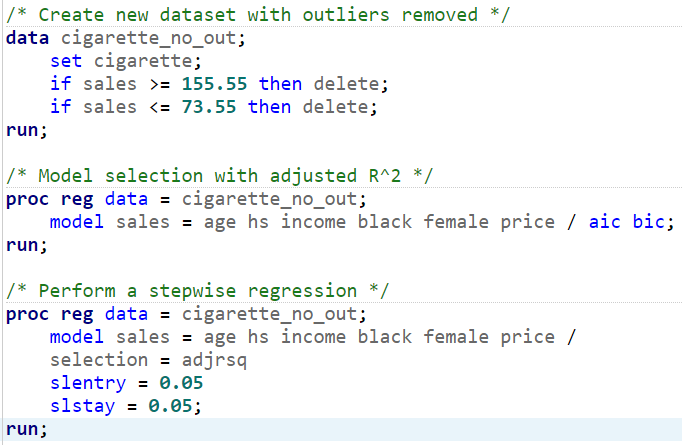
Outliers = Observations > Q3 + 1.5\*IQR or < Q1 – 1.5\*IQR

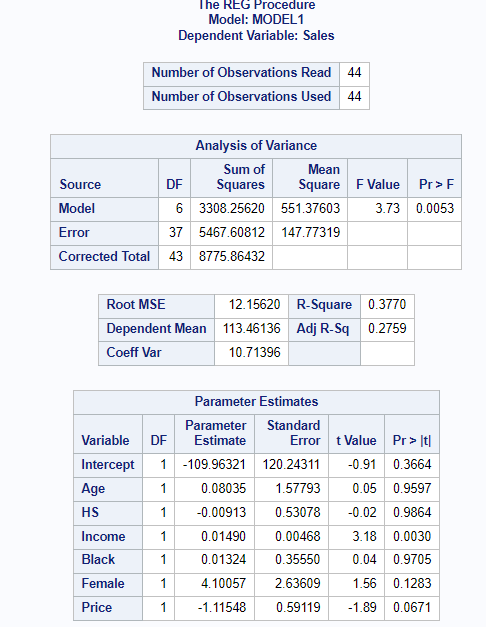
The interquartile range is: Q3 – Q1 = 124.8 – 104.3 = 20.5

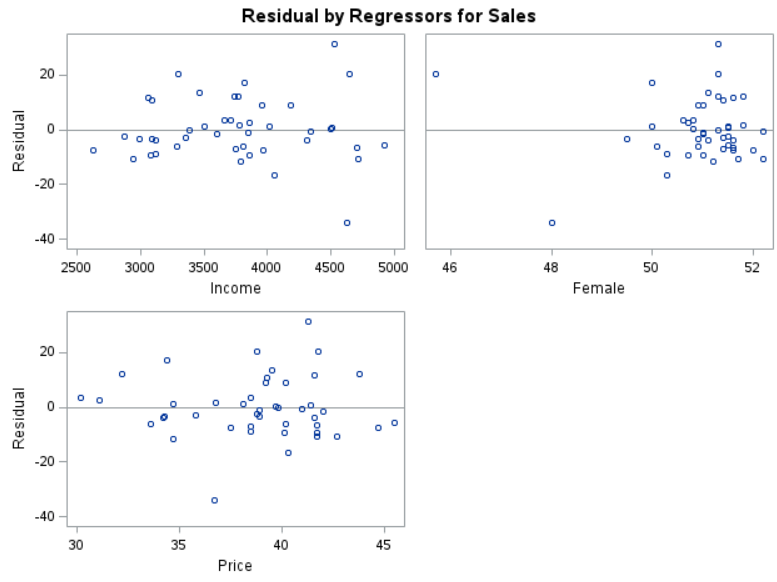
The upper limit for outliers would be: Q3 + 1.5\*IQR = 124.8 + 1.5\*20.5 = 155.55.

The lower limit for outliers would be: Q1 – 1.5\*IQR = 104.3 - 1.5\*20.5 = 73.55.

The outliers are identified as being over 155.55 or under 73.55.

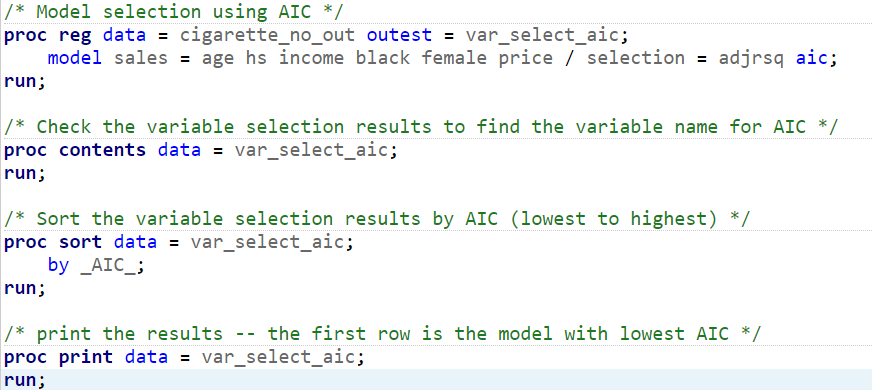




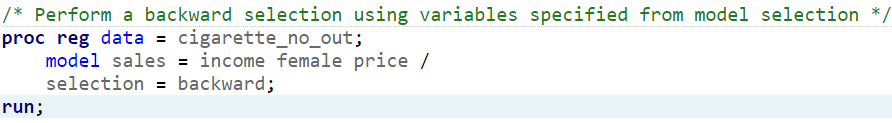


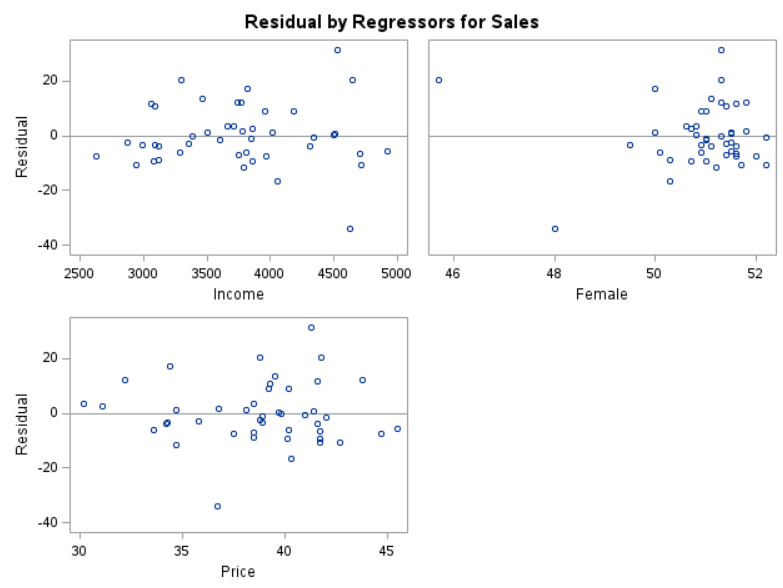
Interestingly, we do not get the same final model. Income and Price remain the same, but we have the inclusion of female rather than age.

* 1. Delete any outlier you find. Redo part (b). Do you obtain the same final model?



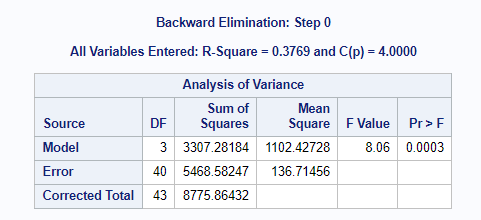






This model also resulted in a change of the final model by switching age with female.

* 1. Use the SSE of the final model in part (e) to numerically calculate the AIC (Hint: refer to lecture note 16 slide 16). Compare your AIC with the AIC output from SAS.



SSE = 5468.5825

AIC = n \* log(SSE/n) + 2(p + 1)

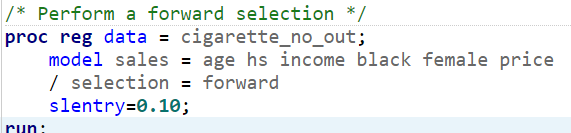
AIC = 44 \* log(5468.5825 / 44) + 2 (4 + 1) = 102.1546

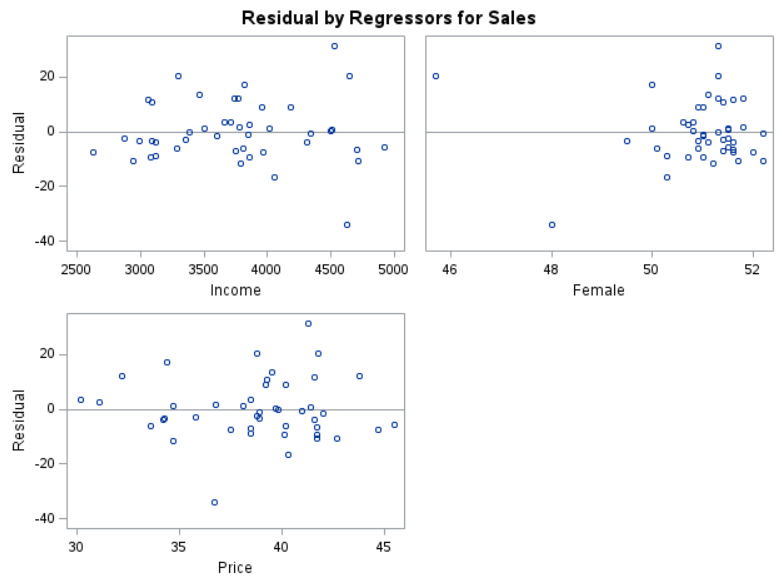
The AIC output from SAS was 220.194. This is over a 72% difference.

* 1. Interpret the regression coefficients of Age and Black in your final model in part (e).

Age and Black were not present in my final model in part e, and therefore, were not determined to be required to fit the model well to the data.

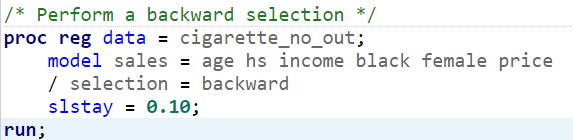
* 1. Delete any outlier you find. Using the forward selection procedure with p-value < 0.10 as the entry criterion, find the final model.

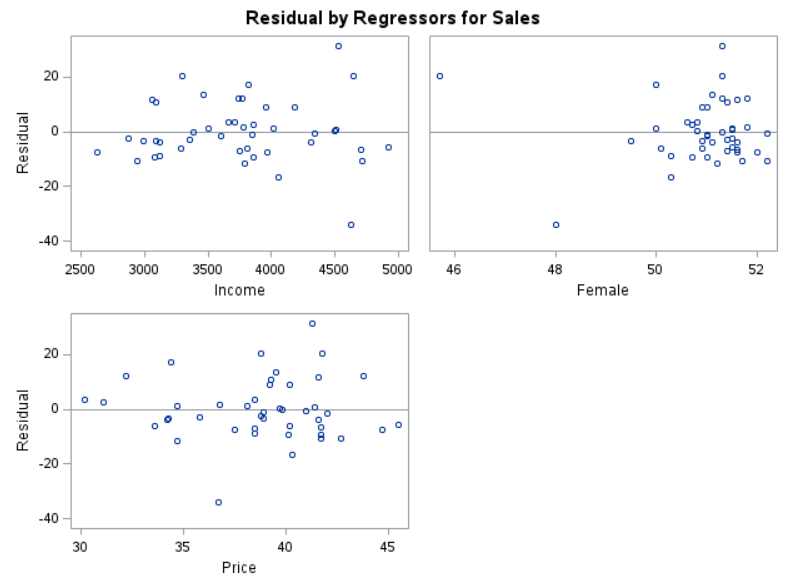




The final model utilizes income, female, and price.

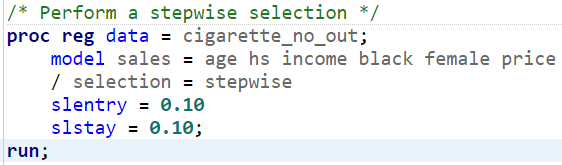
* 1. Delete any outlier you find. Using the backward elimination procedure with p-value < 0.10 as the staying criterion, find the final model.

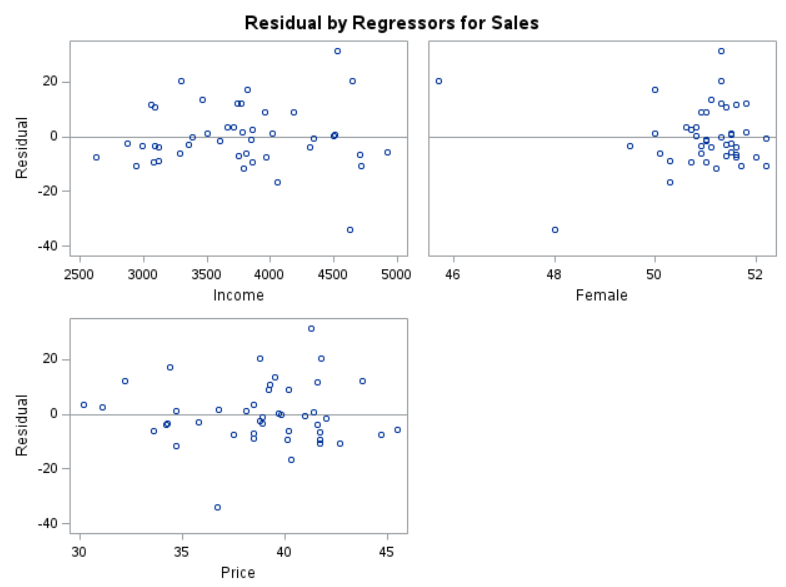




The final model utilizes income, female, and price.

* 1. Delete any outlier you find. Using the stepwise selection procedure with p-value < 0.10 as the entry criterion and p-value < 0.10 as the staying criterion, find the final model.





The final model again uses variables income, female, and price.