Stat 587 section 2 – self assessment for model selection

Scenario: You have collected data on 8 variables measured on 40 subjects. Your goal is to predict Y from some combination of X1, X2, X3, X4, X5, and X6. The data are in modelsel.csv. Notice that there is also an Interest variable that we won’t consider until question 13. Don’t consider transforming any X variables. Context-appropriate transformations have already been done for you.

1) Look at the scatterplot matrix of Y and the 6 X variables (X1, X2, X3, X4, X5, and X6). Any concerns about outliers, influential points, or the assumption that Y and an X are linearly related? Are any X variables strongly correlated with each other?

The first few questions concern diagnostics. Fit a multiple linear regression predicting y from the 6 X variables: X1, X2, X3, X4, X5, and X6. Remember, ignore Interest for now. Use diagnostics from this regression to answer questions 2 – 6.

2) Calculate standardized residuals (= internally studentized residuals on JMP) for each observation. Either look at all 40 values, or plot standardized residuals (Y) vs observation number (X). Do you see any issues?

3) Plot standardized residuals (Y) against raw residuals (X). If you want to identify outliers, do these two types of residuals give you the same answers? What unique feature is present in the standardized residuals is absent from the raw residuals?

4) Calculate Cook’s D for each observation and plot that against the row number. Do you have any concerns about influential observations?

5) Calculate VIF values for each variable. Do you have any concerns about multicollinearity for any of the 6 variables?

6) In question 5), You should have identified 2 variables with very high VIF values. Look back at the scatterplot matrices from question 1). Can you see why those two variables have very high VIF?

You now want to select a subset of variables to predict Y.

7) If you use AIC as your criterion, what is the best set of X variables to predict Y?

8) Using AIC, is the best model substantially better than any alternative, or are there various reasonable alternatives to the best model?

9) If you use BIC as your criterion, what is the best set of X variables to predict Y?

10) Is it reasonable that BIC would select fewer variables than AIC? Briefly explain why or why not.

11) The X5 variable is included in all of the five best models using BIC. Can you conclude that there is no need to transform X5? Explain why or why not.

12) You are considering expanding the set of X variables by including log transformations of X1, X2, X3, X4, X5, and X6. That gives you 12 X variables. Is this is a good idea or not?

Now consider the Interest variable. The goal of the study is to assess the association of Interest and y after adjusting for relevant covariates.

13) Your focus is to understanding the relationship between Y and interest, after adjusting for relevant covariates? If you use AIC with all 7 X variables (X1, X2, X3, X4, X5, X6, and Interest), the best 5 models are:

 model nvar Rsq AdjRsq Cp AIC BIC

 x1 x4 x5 Interest 4 0.8952664 0.8832969 4.725259 -78.56454 -71.80902

 x1 x2 x4 x5 Interest 5 0.8992879 0.8844773 5.391912 -78.13069 -69.68629

 x1 x4 x5 x6 Interest 5 0.8992710 0.8844580 5.397497 -78.12400 -69.67960

x1 x2 x4 x5 x6 Interest 6 0.9034805 0.8859315 6.001829 -77.83151 -67.69824

 x1 x2 x3 x5 Interest 5 0.8966538 0.8814559 6.265252 -77.09796 -68.65357

Do results tell you there is a relationship between interest and Y? Briefly explain why, why not, or whether you need additional information.

Dr. Dixon’s answers:

1) Here’s the plot:



Nothing really, really jumps out at me. There is one point with low values for X3 and X4 that might have some influence on y. X3 and X4 are clearly related to each other. No obvious non-linear relationships between Y and any X variable.

2) Here is the plot of standardized residuals against observation number (next page).

 I don’t see any issues. There are no extremely large (positive or negative) residuals. And, very few of the standardized residuals are < 2 or > 2.



3) Here is my plot:



These two residuals give almost identical information. An observation with a large raw residual is also one with a large standardized residual. However, there is additional information associated with the standardized residuals. They range from ca -2 to ca 2 and they are unitless. The raw residuals range from -7 to 7 for these data and have units. Because the standardized residuals are unitless, a value of -3 is unusual. You can’t interpret raw residuals this way because they have units.

4) Here’s my plot:



The 16’th observation has a D value > 0.2. However, this is considerably less than the usually accepted threshold of 1. So, no concerns with influential observations.

5) VIF values for the 6 X variables are:

|  |
| --- |
|  x1 x2 x3 x4 x5 x6  2.65 2.47 104.53 103.36 1.08 1.10  |

Yes, there are serious concerns with multicollinearity for X3 and X4. Both have VIF values much larger than 10.

6) X3 and X4 are almost perfectly linearly related in the scatterplot from question 1). That’s why have such large VIF values.

7) The model with X1, X2, X4, X5, X6 has the smallest AIC.

8) No, that model is not the only reasonable model. Seven other models have AIC values within 2 of the best.

9) The model with X1, X4, X5 has the smallest BIC.

10) Yes, BIC puts a higher penalty on complexity than does AIC, so when the selected models differ, BIC selects a model with fewer variables.

11) No, inclusion of a variable only says that the linear relationship helps predict Y; it says nothing about whether some non-linear relationships would do better. Another way to make the same point is that inclusion of a variable, even in many “reasonable” models, does not mean there is no lack of fit.

12) This would be a good idea if the data set was sufficiently large. However, I would say no for this data set. My suggested criterion is at least 4 observations per variable “thrown in” to the variable selection, and preferable 10 or more observations per variable. For these data, I would say no, because we now have too many potential X variables (12 X variables with 40 observations). This is less than 4 observations per potential X variable.

13) No. The better approach is to find the best model not including Interest and then estimate the Interest effect. Using AIC, when you add Interest to the model with X1, X2, X4, X5, X6, the estimated slope for Interest is 0.123 (se = 0.080, p = 0.13).

Note: I made up these data, so I know the true model. It was: mean(y) = X1 + 2 X3 + X5 + X6

No effect of interest!

Notice that none of the models discussed above select X3. This isn’t too surprising. Because of the strong correlation between X3 and X4, only one of those two is likely to go in the model. For this specific data set, it turned out to be X4. If I simulated another data set from the same model, it is quite possible that X3 will go in the model, not X4.

Note 2: I have not given AIC or BIC values for any model because each statistical program reports different numbers. But, the difference between two models is the same in every program. One program might report AIC = 2.5 for one model and 2.9 for another. Another program would report -4.7 for that first model and -4.3 for the second. Different numbers, but the same difference. That means the best model (minimum AIC / minimum BIC) is the same for all three programs.

The reason for the different numbers is that there are some constants that could be included in the AIC/BIC computation. SAS / JMP / R make different choices about whether to include (or not) those constants in the computation of AIC / BIC. The constants depend on the number of observations, but don’t depend on which model is being fit.