11.5 This problem uses the data set P313.

a. The answer to this question is no, because many of them would be redundant. This can be stated in more technical terms by noting that there is multicollinearity among the variables, which we have not yet discussed in class. You do not need to answer this part of the problem.

b. The three explanatory variables that they are suggesting are \(X_1\) (local taxes), \(X_6\) (number of rooms), and \(X_8\) (age of the house). Take as a full model \(Y \sim X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9\). As a reduced model, leave only the suggested three explanatory variables in. Compare these two models with an \(F\)-test via the `anova()` function. If the reduced model is not adequate by this test, then your answer is “no” and this part of the problem is done.

If the reduced model is adequate by this test though, you should still look for a better model (even if the reduced model isn’t nested in it). For this, run a `step()` function on the full model and see which model you end up with. If you end up with the reduced model, then your answer is “yes” (the reduced model would adequately describe the sale price) and this part of the problem is done. If you end up with some other model, compare it to the reduced model with an \(F\)-test if the reduced model is nested in it. If the reduced model is not nested in it, compare it to the reduced model with the `AIC()` function instead. Whichever is better by this process will tell you whether your answer should be “yes” or “no”.

c. Repeat the process from the previous part, only with the reduced model as \(Y \sim X_1\) (since \(X_1\) is the local taxes variable). If you disagree with the assertion that this model is adequate, you should present a model that you think is better, explain why (in terms of some statistical test or tests) you think it is better, and discuss what part of the real estate expert’s argument you think doesn’t hold true. If you agree with the assertion, then you should explain why.

11.6 This problem uses the data set P259. The data set itself and the description of the variables is on page 256. Note that the variables are named \(X.1\), \(\ldots\), \(X.9\), \(X.10\), \(X.11\). (yes, three periods on those last two) for some inexplicable reason.

a. As in Problem 11.5, the answer is no because of relationships among the explanatory variables (more specifically, because of multicollinearity). You do not need to do this part of the problem.

b. (Do you like how they have (a)-(f) all within Part (b)?) Since these are not nested, compare them using the AIC. This should tell you which one of the six to choose. Oh, and in spite of the strange wording in (f), all that is meant is to use \(X_5\), \(X_8\), and \(X_{10}\) as explanatory variables. Also recall that the `AIC()` function can have as many arguments as you like (separated by commas). As for finding a better model, run a `step()` on the full model consisting of all 11 explanatory variables (without interaction terms). If that gives you the “best” of the six models, then you needn’t search further. (Just answer “no”, that you don’t see a better model.) If that gives you a different model, then compare that to the “best” of the six with `AIC()` and use the better one as the “best” model that you found.
c. You should do the plots that they recommend and compare them to the functions depicted on page 154. However, do not include your plots or anything written for this part of the problem. (Just note to yourself what you see in the plots.)

d. As with the previous part, you should do the plots for yourself (and they should look more linear), but do not turn anything in for this part of the problem.

e. As it says, repeat Part (b) using $W$ in place of $Y$. By “what are your conclusions”, they are asking how your results in this section compare to those in Part (b). For example, did you arrive at the same “best” model? If not, which of the two is better and how do you know? (Remember that you can compare models with the AIC even if they aren’t nested.)

f. To “regress”, set up the linear model in question and output its summary.

g. You don’t need to write a “brief report”, but instead just compare the model from Part (f) to your previous “best” model (using the AIC). Based on this comparison, state what the “best” overall model (at least by AIC standards) that you have found in this problem is. Notice that in addition to comparing many non-nested models, you have also tried some transformations of variables based on scatterplots in your search for the “best” model. Those are both quite typical in the model-selection process.

Now throw a party — you have just finished your last Math 260 homework!