STAT 325/425: Data Analysis and Linear Models

Homework 8 (and 9)

Due: Tuesday, December 04, 2018

This last homework is a case study of the Real Estate Sales data in Appendix C.7 of the textbook[1].

The Case study:

Residential sales that occurred during the year 2002 were available from a city in the midwest. Data on 522 arms-length transactions include sales price, style, finished square feet, number of bedrooms, whether the house has a pool, lot size, year built, whether air conditioning is installed and whether or not the lot is adjacent to a highway. The city tax assessor was interested in predicting sales price based on these demographic variable information. Your job is to build a model that can do this.

The assignment is to write a report with Section headings.

1. **Introduction**
2. **Exploratory Data Analysis**
3. **Model Building**
4. **Model Diagnostics**
5. **Model Validation**
6. **Conclusion**

More details of what these sections should contain are given below

General Instructions

- This report will count as two homework assignments (worth 40 points).
- *Type* your homework and write full sentences about your findings. Be concise and to the point.
- You may choose whether you want to do this homework using Rmarkdown.
- Your `R` code should be included in an Appendix, *not* in the main text of your report.
- Tables of results should be in the main text but can be a copy of the `R` output.
- Number your Figures (and Tables if you have them) and refer to these numbers when you are discussing what they show. Only include Figures that are relevant to what you want to say.
- **Page limit:** Your report (not counting the appendix) may not be more than 12 pages.

[1] This case study is based on Problems 9.31, 9.33 and 10.31 in the book

Specific instructions for each Section

Begin by selecting a *random* sample of 300 observations to use in your model-building effort. The remaining 222 will be used for validation.

Hint: Store the test dataset separately or use `set.seed(<pick a number>)` so that you don't end up working with several different testing datasets.

1. **Introduction:**
   Introduce the project.

2. **Exploratory data analysis:**
   Perform exploratory data analysis. Your goal here is to identify the general form of a regression model, i.e. identify whether transformations are needed of the response and/or predictors. Include at least the following:

   (a) Scatterplot matrix for the quantitative variables. What does this figure tell you about the (quantitative) data?

   (b) Fit a first-order regression model to the "raw" test data. That is, regress sales price on all 11 predictors. Provide the standard diagnostics plot (i.e. the four figures you get from `plot(Fit)`). Are model assumptions violated? - If so, try again with a transformation. No more intensive model checking is required at this stage.

   (c) Decide the pool of predictor variables:

      i. If you decide to transform the response provide scatterplot of the transformed response against the quantitative predictors.

      ii. Plot the (possibly transformed) response against qualitative variables in the form of side-by-side boxplots.

      iii. Provide an *Added-variable plot* for the following variables: square feet, number of bedrooms, number of bathrooms, garage size, year and lot size.

      iv. Based on the figures you provided in (i)-(iii) decide what pool of predictor variables you will use in the model building step and in what functional form (e.g. linear term, quadratic term etc.)

   Helpful hints:

   - Air conditioning, pool and highway are qualitative variables that have only two categories and are coded as 0 and 1 so you can use them as they are.
   - Other qualitative variables are Quality (3 categories) and Style (10 categories). Use `factor()` in the formula for `lm()` to let R make the indicator variables for you. This will also make sure that R treats these indicator variables as a group (useful for the model building step).

- **Added-variable plots**: You can either do them "by hand" by fitting the required regression models and plot residuals. Another way is to install the `car` package in R and use the function `avPlots()`, e.g.

  ```
  > avPlots(FitAll, terms = ~ year)
  ```

3. **Model building:**
   Using the group of predictors and the (possibly transformed) response see if you can reduce the number of explanatory variables. Your analysis should include the following

   (a) Order all possible subsets based on the AIC, BIC and R_a^2 criteria. Plot the criteria for all sub-models versus number of variables. Are the criteria optimized for models of different sizes?

   (b) List the best five models picked by each criteria. Are these models generally the same or similar? If not, in what way are they different? Are some predictors consistently present in all the "best" sub-models?

   (c) Using the `step()` function in R do stepwise regression (both directions). Is the model identified by this procedure the same as any of the one identified in part (b)?

   (d) Pick two or three models to continue with in the next part. Try to pick at least one model that is more parsimonious than the others (note: it is likely that BIC picks more parsimonious models than AIC does)

   Hints

   - You can use the R code provided in lecture as a template for this. There may be R packages that can do this easier but I did not find one that fits for this problem. You have to make sure that all the indicator variables for one qualitative variable are treated as a group (all included or all removed together) and if you have polynomials the lower order terms should always be included when higher order terms are included.

4. **Model Diagnostics:**
   Do more extensive model diagnostics on the two or three models you picked in part 2. This should include for each model:

   (a) The "traditional" residual plots, e.g. check if variance is constant, check normality etc.

   (b) Obtain the studentized deleted residuals (denoted t_i in the book) and plot them against the fitted values. Do these tell a different story than the semi-studentized residuals (denoted e_i^* in the book)?

(c) Summaries of each model. Are all predictors marginally significant? Are the R^2 very different for these models?

(d) Identify influential observations by calculating Cook's distance for each observation and plot them (against the case number). Are there influential observations? Which ones?

(e) Delete influential observations (if any) and refit the models before doing model validation.

Hints:

- To obtain the diagonal of the hat matrix (the h_{ii} values) you can use the R-function `hatvalues(Fit)`
- R can calculate Cook's distance for you: `cooks.distance(Fit)`

5. **Model Validation:**
Do model validation of the two or three models you picked in part 2. If influential observations were identified in part 3, use the model fits where they have been removed. This analysis should include for each model:

(a) Predict the validation dataset and obtain the mean squared prediction error (MSPR) and compare to the estimated MSE. What does this comparison tell you about the adequacy of the model?

(b) Fit the model to the validation dataset and compare estimated regression coefficients (through confidence intervals), MSE and R^2. How do these compare?

6. **Conclusion:**
Pick *one* model to present to the city tax assessor. Use the foregoing analyses to justify your choice. Fit the selected model to all the data, i.e. both the test data and validation data (you may exclude any influential observations), and present that model fit as you would to the city tax assessor. For example, interpret some of the regression coefficients, which ones you believe are most important in determining sales price and how much of the variation in sales price you were able to explain.