Homework 8

36-401, Fall 2015, Section B

Due at 4:30pm on Friday, Nov 13, 2015

This is a practice project for multivariate data analysis. Write up your work in a document of **at most 7 printed pages**. This should answer all the prompts and specific questions below, but also read as a single connected report.

Format: The 7 page limit includes **all text and figures**. Code should either be integrated with R Markdown, and hidden, or go in a separate appendix. If you use R Markdown, submit both the .Rmd file and the knitted PDF. If you do not, submit a PDF report and a *separate* R appendix.

Research Scenario: Predicting house prices People buying or selling houses would like to know how much they can expect to get, or pay, for a property. This is also a concern for those who are making mortgage loans, or for those taxing real estate (and who are more likely to commission statistical studies than individual home-owners). The price of a house depends on its physical characteristics, including size, features, quality of construction, age, etc. It also depends on location, and current market characteristics. You are approached by a research group which has a data on a sample of residential sales in a midwestern city; the variables are described in Table 1. They would like you to fit a multiple linear regression, with sales price as the response and the other variables as predictors.

Your client believes that higher quality of construction should predict higher prices. They also believe that older houses tend to have lower prices, though this relationship is thought to differ depending on whether or not the house is adjacent to a highway. They also think that the relationship between price and finished area differs depending on the number of bedrooms.

Suggested Outline

1. *Introduction* Write four to five sentences introducing the research problem and describing the specific research hypothesis. Cite any information sources in parentheses.

2. *EDA* How many observations do you have?

    - Examine the (predictor and response) variables univariately and multivariately.

Variable Name	Description
Sales price	Sales price of residence (dollars)
Finished square feet	Finished area of residence (square feet)
Number of bedrooms	Total number of bedrooms in residence
Number of bathrooms	Total number of bathrooms in residence
Air conditioning	Presence or absence of air conditioning: 1 if yes; 0 otherwise
Garage size	Number of cars that garage will hold
Pool	Presence or absence of swimming pool: 1 if yes; 0 otherwise
Year built	Year property was originally constructed
Quality	1= high quality, 2 = medium, 3 = low
Lot size	Lot size (square feet)
Adjacent to highway	1 if the property is adjacent to a highway, 0 otherwise

Table 1: *Variables in the data set*

- Provide graphical displays/numerical measures for all variables.
- You need EDA for all pairs of continuous variables and at least for the categorical variables and the response variable. Describe your results.
- Which variables seem associated with the sales price?

3. *Initial modeling* Start by building a multivariate linear regression to the data predicting the sales price from the predictor variables. Address the following when building that model:
   - Variables like number of bedrooms/bathrooms, construction quality, and garage size could be coded multiple ways: continuous, nominal, ordinal. Justify your choices.
   - A colleague hypothesizes that there might be an interaction between the finished square feet of the house and the number of bedrooms. Use graphs to generate evidence for or against this hypothesis. Do you agree? Try the new interaction term in your model. Do you keep the interaction in your model? Why or why not?
   - Similarly explore an interaction between the year when the house was built and whether or not the house is adjacent to highways.
4. *Diagnostics/model selection*
   - Are the basic assumptions met for your multivariate linear regression model? Why or why not?
   - What transformations do you choose (if any)? Why?
   - Are there any outliers in your sample overly influencing your model? Identify any outlier candidates and decide whether or not to remove them. Give details.

- Do you exclude any variables? Why? All exclusions/inclusions must be justified.

5. *Final model inference/results* Create a table that summarizes your final model (coefficients, standard errors, confidence intervals, p-values). Provide interpretations of all your coefficients in the context of the problem. Be sure to address the specific questions of the client:

   - whether older houses have lower sales prices;
   - whether the relationship between age and sales price depends on adjacency to a highway;
   - whether higher construction quality have higher sales prices;
   - whether the relationship between finished area and sales price varies with the number of bedrooms.

6. *Discussions/results*: What are your conclusions? Identify a few key findings, and discuss, with reference to the supporting evidence. Can you come up with explanations for the patterns you have found? Suggestions or recommendations for the client? How could your analysis be improved? (6–8 sentences)

Rubric

Words (10) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (5) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (5) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text.

Code (10) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. The text of the report is free of intrusive blocks of code. If you use R Markdown, all calculations are actually done in the file as it knits, and only relevant results

are shown[1]. If you do not use R Markdown, the code in your appendix must generate exactly the results you show in your report, and must have comments making it clear which parts of your code go with which results.

Explotory Data Ananlysis (15) Variables are examined individually and bivariately. Features/observations are discussed with appropriate figure or tables. The relevance of the EDA to the modeling is clearly explained.

Model formulation and checking (30) The initial model's formulation is clearly related to the substantive questions of interest. The model's assumptions are checked by means of appropriate diagnostic plots or formal tests; if the model is re-formulated, the changes are both well-motivated by the diagnostics, and still allow the model to answer the original substantive question. Limitations from un-fixable problems are clearly noted.

Estimation, Inference and Uncertainty (15) The actual estimation of model parameters or predictions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty.

Conclusions (10) The substantive questions about real estate pricing are answered as precisely as the data and the model allow. The chain of reasoning from estimation results about models, or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.

Extra credit (10) Up to ten points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.

Writing Advice

Your language should be very clear and precise. Do not make claims for which you have no evidence. Do not say "will" or "would" when you really mean "may" or "might". Do not use language that implies causation; you are studying associations between variables only. Move away from wordy phrases (e.g: This is because, this is due to, the reason that this is, this means that, I believe that this, I think the reason is that). Make sure pronouns have clear referents ("these results show" vs. "this shows").

[1] See the model report for DAP 1 for examples of how to do this.