Journal of Statistics Education

ISSN: (Print) 1069-1898 (Online) Journal homepage: https://www.tandfonline.com/loi/ujse20

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock

To cite this article: Dean De Cock (2011) Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Journal of Statistics Education, 19:3, , DOI: 10.1080/10691898.2011.11889627

To link to this article: https://doi.org/10.1080/10691898.2011.11889627

Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project

Dean De Cock
Truman State University

Journal of Statistics Education Volume 19, Number 3(2011),
www.amstat.org/publications/jse/v19n3/decock.pdf

Key Words: Multiple Regression; Linear Models; Assessed Value; Group Project.

Abstract

This paper presents a data set describing the sale of individual residential property in Ames, Iowa from 2006 to 2010. The data set contains 2930 observations and a large number of explanatory variables (23 nominal, 23 ordinal, 14 discrete, and 20 continuous) involved in assessing home values. I will discuss my previous use of the Boston Housing Data Set and I will suggest methods for incorporating this new data set as a final project in an undergraduate regression course.

1. Introduction

My first exposure to the Boston Housing Data Set (Harrison and Rubinfeld 1978) came as a first year master's student at Iowa State University. Its analysis was the final assignment at the conclusion of the regression segment within our statistical methods class. The assignment was fairly open ended with a brief description of the data set and the simple task of finding a good model for the prediction of housing prices. At the time, the data set seemed similar to others I had encountered and it slipped from my memory until seven years later when I found myself as a new faculty member teaching my first regression course. Although I had only recently begun my career in academia, I had already established the desire to incorporate some of the principles that would officially be recommended in the GAISE guidelines, such as the use of active learning, real data, and group work. In each of my other statistics classes, I had incorporated a final group project that integrated the concepts learned throughout the semester and I wanted to do the same in my regression course.

For a regression project, I was looking for a data set that would allow students the opportunity to display the skills they had learned within the class. The ideal data set needed to have a reasonably large number of variables and observations so that students would have to go beyond a simple algorithm, such as forward or stepwise selection, to construct a final model. At the time, I remembered the assignment from my own past and I searched the web to see if I could find the Boston Housing Data Set (http://lib.stat.cmu.edu/datasets/boston). I was surprised at the number of references and uses of the data set within the academic community and determined that its 506 observations and 14 variables would serve my purposes well. Over the years I have continued to use this data set, but with each passing year I have become more dissatisfied with its use. The original data set is from the 70's and the housing prices have become unrealistic for today's market. I had contemplated inflating the prices by some set amount or scaling factor to obtain more contemporary values but that would change the data from real to realistic, which was not my preference.

As part of my sabbatical leave, one of my goals was to find a new data set that I could use as my final project. Although open to new subject areas, my hope was to find a more recent housing data set as students are typically familiar with the variables associated with home evaluation. I began my search by scouring sites such as DASL and the JSE Data Archive and although I found several potential data sets (e.g. Woodard and Leone 2008), the data sets were rather limited in the number of observations ($n \leq 100$). A chance visit to my alma mater opened the door for the data set presented in this article. In chatting with some members of the Iowa State StatCom group about their current projects, a student mentioned the group was updating the assessment model used by the Ames City Assessor's Office. They described the large number of variables and observations within the data set and I immediately set up an appointment with the City Assessor's Office to discuss the use of the data. After a brief meeting with the Assessor and Deputy Assessor outlining the data and the assessment process, I was given access to the data.

The data came to me directly from the Assessor's Office in the form of a data dump from their records system. The initial Excel file contained 113 variables describing 3970 property sales that had occurred in Ames, Iowa between 2006 and 2010. The variables were a mix of nominal, ordinal, continuous, and discrete variables used in calculation of assessed values and included physical property measurements in addition to computation variables used in the city's assessment process. For my purposes, a "layman's" data set that could be easily understood by users at all levels was desirable; so I began my project by removing any variables that required special knowledge or previous calculations for their use. Most of these deleted variables were related to weighting and adjustment factors used in the city's current modeling system.

2. The Ames Housing Data

After removal of these extraneous variables, 80 variables remained that were directly related to property sales. Although too vast to describe here individually (see the documentation file http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt), I will say that the 80 variables focus on the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property (e.g. When was it built? How big is the lot? How many square

feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?).

In general the 20 continuous variables relate to various area dimensions for each observation. In addition to the typical lot size and total dwelling square footage found on most common home listings, other more specific variables are quantified in the data set. Area measurements on the basement, main living area, and even porches are broken down into individual categories based on quality and type. The large number of continuous variables in this data set should give students many opportunities to differentiate themselves as they consider various methods of using and combining the variables.

The 14 discrete variables typically quantify the number of items occurring within the house. Most are specifically focused on the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. Additionally, the garage capacity and construction/remodeling dates are also recorded.

There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. They range from 2 to 28 classes with the smallest being *STREET* (gravel or paved) and the largest being *NEIGHBORHOOD* (areas within the Ames city limits). The nominal variables typically identify various types of dwellings, garages, materials, and environmental conditions while the ordinal variables typically rate various items within the property. The coding within the original data typically utilized an eight-character name that was relevant to the classification but some of the original class levels were difficult to interpret. For ease of use many class levels were recoded into slightly more usable forms (see the documentation file http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt).

> ***Helpful Hint: Depending on the level of student, instructors may want to decide how much advice/direction they would like to give the students. They may want to code categorical variables into dummy variables ahead of time or may want to give students hints about how to combine or use the available variables. For my purposes I give the students the data "as is" and expect them to determine how the data could best be utilized.***

There are two variables (*PID* and *NEIGHBORHOOD*) that may be of special interest to users of the data set. *PID* is the Parcel Identification Number assigned to each property within the Ames Assessor's system. This number can be used in conjunction with the Assessor's Office (http://www.cityofames.org/assessor/) or Beacon (http://beacon.schneidercorp.com/) websites to directly view the records of a particular observation. The typical record will indicate the values for characteristics commonly quoted on most home flyers and will include a picture of the property. The *NEIGHBORHOOD* variable, typically of little interest other than to model the location effect, may be of more relevance when used with the map (http://www.amstat.org/publications/jse/v19n3/decock/AmesResidential.pdf) included in the supplementary materials.

As a final note, the original data set (n=3970) contained all sales that had occurred within Ames from 2006 to 2010, including stand-alone garages, condos, and storage areas. As these special

sales created unusual conditions (observations with no living space or lot size) I chose to include only residential sales within the data set presented here. Additionally, approximately 100 homes changed ownership multiple times during the 4-year time period. As this gave a greater weight to these particular homes, I elected to keep only the most recent sales data on any property. Removing the multiple and non-residential observations resulted in the final data set containing 2930 observations. The data set is available at: http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls. A text file version of the data set is available at: http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.txt. The documentation file explaining details of the data set is available at: http://www.amstat.org/publications/jse/v19n3/decock/DataDocumentation.txt).

> ***Potential Pitfalls (Outliers): Although all known errors were corrected in the data, no observations have been removed due to unusual values and all final residential sales from the initial data set are included in the data presented with this article. There are five observations that an instructor may wish to remove from the data set before giving it to students (a plot of SALE PRICE versus GR LIV AREA will quickly indicate these points). Three of them are true outliers (Partial Sales that likely don't represent actual market values) and two of them are simply unusual sales (very large houses priced relatively appropriately). I would recommend removing any houses with more than 4000 square feet from the data set (which eliminates these five unusual observations) before assigning it to students.***

3. Using the Data: End of the Semester Project

The regression course at my institution is a one semester course focusing primarily on regression methods with some time series analysis. The course has an introductory statistics pre-requisite (but no matrix or calculus requirement) and is mainly composed of business and psychology majors. Although the project could be assigned at any time during the semester, I prefer to distribute the information after the students have covered the material necessary to complete the project (concepts of multiple regression, assumption validation, and model selection techniques). This desire creates a conflict in that I must cover all the relevant material in a manner timely enough for students to finish the project by the end of the semester. To alleviate problems, I defer all the time series material (distributed throughout our text) until the end of the semester as this material is unnecessary for the analysis of the project. This typically allows about three weeks for the students to complete all components of the project.

The original project I gave my students was very similar to the homework assignment I received as a graduate student in that students were simply asked to use everything they had learned in the class to construct the best model possible. Over the years I have made changes, due mostly to difficulties that have arisen, that I believe have improved both the project and the experience for the students. The most critical of these changes is that the project now comprises three distinct components, each with its own due date. The first component, due one week into the project, requires students to submit a simplistic model (MODEL1) that can be used for predicting the sale price of houses. This component is used to verify that students understand the assignment and are familiar with the methodology for submitting their models. The second component, due two

weeks into the project, requires students to submit a more complex model (MODEL2) that represents their best effort at predicting housing prices. This component will be applied to a validation set to determine a "fit" grade that comprises 30% of their project grade. The final component, due on the last day of class, is a written report that contains all the analysis, interpretation, and information for the two submitted models. The written report completes the remaining 70% of the total project grade.

MODEL1 (the simplistic model) is the focus of the written report and its analysis is intended to showcase the knowledge students have gained with regards to proper techniques for statistical inference. I limit the size and type of model that can be used so that students may focus their attention on the interpretation of the model (rather than becoming mired in the pursuit of finding the "best" model). A strong analysis should include the interpretation of the various coefficients, statistics, and plots associated with their model and the verification of any necessary assumptions.

> ***Helpful Hint: I have found it important to set both a lower and an upper bound on the types of models students can create for MODEL 1 (see Figure 1 for excerpt from my handout). I have learned through experience that without these guidelines students have a tendency to either create models too complex to interpret and analyze or too simplistic to fully demonstrate the modeling skills they have learned in the class.***

MODEL1 – In the first model you are allowed only limited manipulations of the original data set. You are allowed to take power transformations of the original variables [square roots, logs, inverses, squares, etc.] but you are NOT allowed to create interaction variables. This means that a variable may only be used once in an equation [if you use x^2 don't use x]. Additionally, you may eliminate any data points you deem unfit. This model should have a minimum r-square of 73% and contain at least 6 variables. The intent of this project is for the majority of your effort to be devoted to creating and reviewing this model.

Figure 1: Excerpt from project handout (Boston Data)

MODEL2 (the complex model) is intended to allow students the opportunity to construct a best fitting model for predicting housing prices. Students are encouraged to experiment with any of the methods that were discussed during the semester for finding better models and are allowed to create any new variables they desire (such as quadratic, interaction, or indicator variables). MODEL2 is evaluated through a cross-validation or data splitting technique where the original data set is split into two data sets: the training set and the validation set. The students are given the training set for the purpose of developing their model and I retain the validation set for use in evaluating their model. A relative grade is assigned by comparing their fit on the validation set to that of their fellow students with bonus points awarded to those who substantially exceed their fellow students and point reductions occurring for models which fit exceedingly poorly (See section 4 - Evaluating the Models for more details).

> ***Helpful Hint: Although in the past I have given students free reign to create the best model possible for MODEL 2, in the future I plan to set an upper limit on the number of terms a model may contain. With the Boston Housing Data Set students would occasionally submit excessively large models (up to 27 variables including interaction***

terms and polynomials) and I believe this might be exacerbated with the new larger data set.

Although there are no exact rules for the size of the training and validation data sets (Neter, Kutner, Nachtsheim and Wasserman 1996), it is common practice to split the original data set into two equal components. For my purposes, I selected a set of 100 houses for my Boston validation set (80/20 split) which I felt allowed me enough data to evaluate their models while still giving the students a large data set to work with. The two data sets can be easily created by randomizing the original data and selecting the relevant proportion for each component with the only real requirement being that the number of observations in the training set be six to ten times the number of variables (hence my desire to locate a large data set).

Helpful Hint (training/validation): I chose to use randomization to create my Boston sets but those wishing to achieve a more consistent split may want to use a systematic sampling scheme. Simply order the original data set by a variable of interest (such as sale price) and select every k^{th} observation to achieve the desired sample size (k=2 for a 50/50 split or k=4 for a 75/25 split). While not necessary, I chose to create new training and validation sets each semester to prevent students from directly using reports from previous terms to complete their project.

I have found it very important to require the students to test their models with a single hand calculation (see Figure 2 for excerpt from my handout). Students have a tendency to blindly accept whatever model is output from their software and it never occurs to them that they should check their answer. Although students find hand calculating the fitted value for a data point redundant (they feel they can simply copy the fitted value from the output), it will often expose errors in their model construction. The most common error I have found is students losing track of what they have done in creating complex variables such as transformations and interactions (i.e. they think that their new variable v13 is an interaction between v1 and v3 when in actuality it is some other combination or transformation).

Model Check - Also test your equation on the first observation in the data set to make sure that the model gives a reasonable answer and include this on the last page of your report. This should be done BY HAND using a calculator (this means use the raw data from the original spreadsheet and manually calculate all transformations and interactions with your calculator)! Models that do not give reasonable answers will be given a minimum 2 letter grade reduction. Also be careful as you cannot use certain transformations [ln or inverse x] if a variable has values of 0.

Sample Submittal (please do not use the equation editor)
Yhat = -27810.661 + 9322.460(X1) - 4798.086(SQRT(X5)) + 5119.678(LnX10) - 110.087(X1X10)
Yhat = -27810.661 + 9322.460(6.575) - 4798.086(2.32) + 5119.678(0) - 110.087(6.575*1) = 21629
Predicted value = $21629
Actual value = $23999
Model seems reasonable (Error = $2370)

*** NOTE – it is very important that you have at least three significant figures in each of your coefficients!!! So if you have a value like 0.002X1 you need to convert it to 0.00195X1 ***

Figure 2: Excerpt from project handout (Boston Data)

Alternative Application: My purpose in the formation of this data set was to create a rich example where introductory students could focus on modeling rather than on data preparation. I can see value in giving the original data set to more advanced students and letting them experience the problems with cleaning "raw" data. As such I am willing to share the original data set, as received from the Ames Assessor's Office, with any interested parties.

4. Evaluating the Models – Class Discussion

I spend the last day of the semester discussing the student models in the classroom. Although students tend to be less than enthusiastic about new material on this day, I have found they are very interested in reviewing how their model performed relative to others in the class. I begin by having a member of each group step up to the board and write out their final model. Although I ask them to bring a copy with them, I find it a good idea to print them myself before class as there always seems to be at least one group who forgets. We begin our review by looking over the models and discussing general trends. Typically the models are very similar but some groups will have done unusual things (transformations, interactions, etc.) that might be worthy of special attention and discussion. I believe that students enjoy seeing what other groups did as many of the groups are somewhat "secretive" during the modeling process as they are trying to create a better model than their peers.

Following the qualitative discussion of their models, I transition into the quantitative evaluation of the fit of their models. I remind the students of the concept of the validation set (mentioned earlier in the semester) and then talk about the four main criteria I use for evaluating their model. In each measure, the actual home price (Y) of each observation in the validation set is compared the predicted value (Yhat) obtained from their model.

- **Bias** – *Average (Yhat-Y)* – This concept is the easiest for the students to understand as positive values indicate the model tends to overestimate price (on average) while negative values indicate the model tends to underestimate price.

- **Maximum Deviation** - *Max* |*Y-Yhat*| - Students also find this measure easy to understand as it identifies the worst prediction they made in the validation data set.

- **Mean Absolute Deviation** – *Average* |*Y-Yhat*| - Although not as intuitive to the students, once contrasted with bias, students grasp that it is the average error (regardless of sign).

- **Mean Square Error** – *Average* $(Y\text{-}Yhat)^2$ – The least intuitive and least meaningful measure for the students. I only include it so that I can compare its calculation to the methodology used to obtain the coefficient estimates from the original data set (linking back to the idea of Least Squares Regression).

I present the results of the four quantitative measures in graphical form (Figure 3) so students can see how they performed relative to their peers. Commonly the models from most groups will

perform approximately the same with only a few groups setting themselves apart as being much better or worse than their fellow students (such as Group 11's excellent bias result or Group 4 & 7's poor bias, mean absolute error, and mean square error performance on the 2007 data).

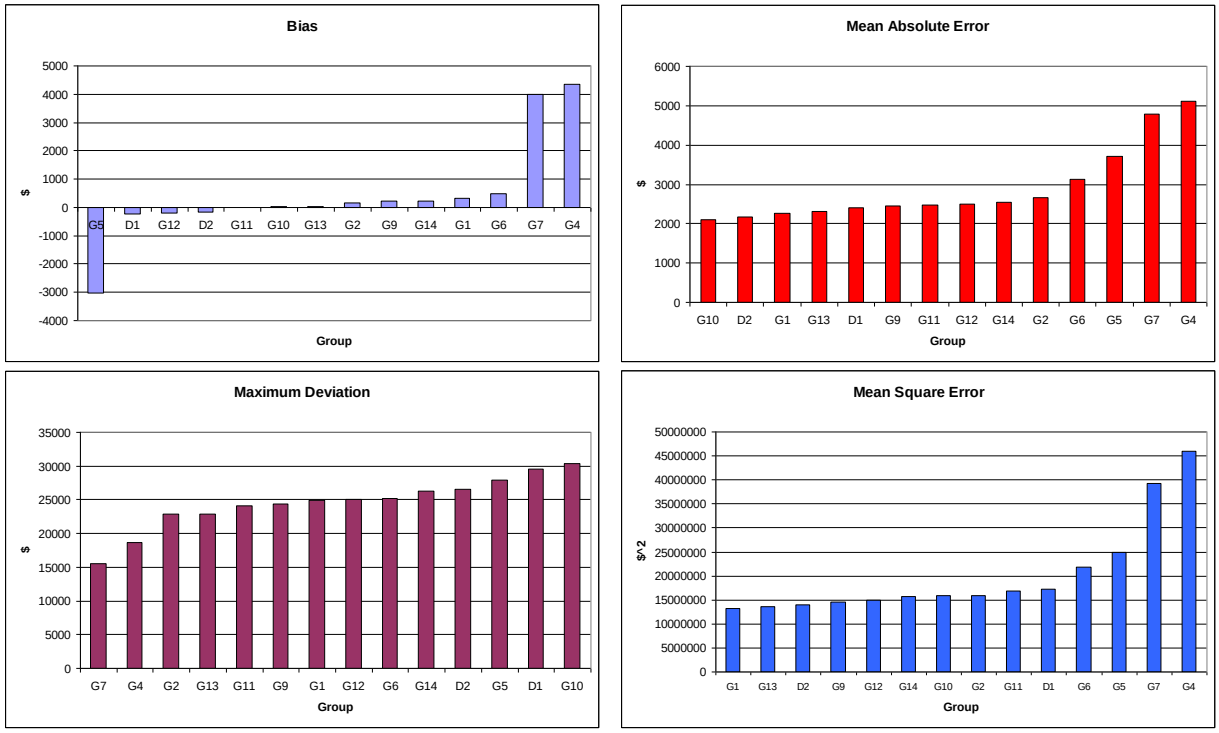

Figure 3: Graphical summary of measures for evaluating model fit (Spring 2007; Boston Data)

> ***Helpful Hint: I often create my own versions of MODEL1 & 2 from the training data set to use as benchmarks for gauging student performance. I hold myself to the same criteria as the students for constructing these models; with the added constraint of only allowing myself a very short time period for their construction. My performance (models D1 & D2) on the validation set can be seen in Figure 3 and you will notice that several groups outperformed me in each of the categories.***

5. Adjustments for Lower-level Classes

Although the data set could be used in its current form for any level statistics class, I would recommend several changes to facilitate its use with lower level students. My main criteria for a good project data set was a large number of variables and observations, but introductory students may find themselves overwhelmed by the amount of information in this data set. The first modification I would recommend is reducing the number of observations within the data set. My personal preference is to create a data set that will allow histograms and dotplots to show general shapes and trends while simultaneously having few enough observations so that individual points can still be seen on scatterplots. Although this will depend on the specific plot or activity, about 200 data points will likely suffice. While a simple random sample from the original data could be used to draw the appropriate sized subset, I will make a few more specific recommendations.

- **Eliminate all sales except for the "normal" from the *SALES CONDITION* variable**. Unless an instructor specifically wants to create an activity that investigates the difference between the various types of sales (foreclosures, new homes, family sales, etc.) the different conditions will simply serve to complicate the results and confuse the students.

- **Remove all homes with a living area (*GR LIVE AREA*) above 1500 square feet.** The purpose to the second step is to alleviate problems with non-homogeneous variance. As might be expected there is increasing variation with increasing price within the Ames housing market. This problem can be remedied by taking a transformation (square root) of the sales price but those wishing to keep the response in dollars can simply use the smaller homes as they tend to show more homogeneous variation.

- **Select desired observations.**

  Although the data set could be reduced further using other variables (such as using only one story homes or eliminating homes without a basement), I would recommend drawing a simple random (or systematic) sample from the remaining "normal" sales to create the final data set.

  Helpful Hint: While higher level programs (such as SAS and R) can easily cull the appropriate observations, the same results can be achieved in Excel by simply creating a column of random digits adjacent to the existing data set (using the RAND function or Data Analysis Tool) and sorting the data by this or other columns. After sorting, the appropriate size sample can simply be cut and pasted from the original data to create the reduced data sets. (Note users of Minitab will find the* Calc-Make Patterned Data – Simple Set of Numbers *and the* Data-Subset Worksheet *commands useful for creating smaller data sets.)

After reducing the number of observations, the instructor is free to use as many of the variables as they wish in their activities. Most introductory activities will likely want to focus on simply using one or more of the available continuous variables. The most obvious simple regression model is to predict sales price based on above ground living space (*GR LIVE AREA*) or total square footage (*TOTAL BSMT SF + GR LIV AREA*). The total square footage model (Figure 4) indicates some possible curvature which could lead into discussions of quadratic variables. Instructors wanting to avoid this solution may want to review the Pardoe (2008) paper which sites a (non-quadratic) method used in real estate to deal with non-linear relationships.

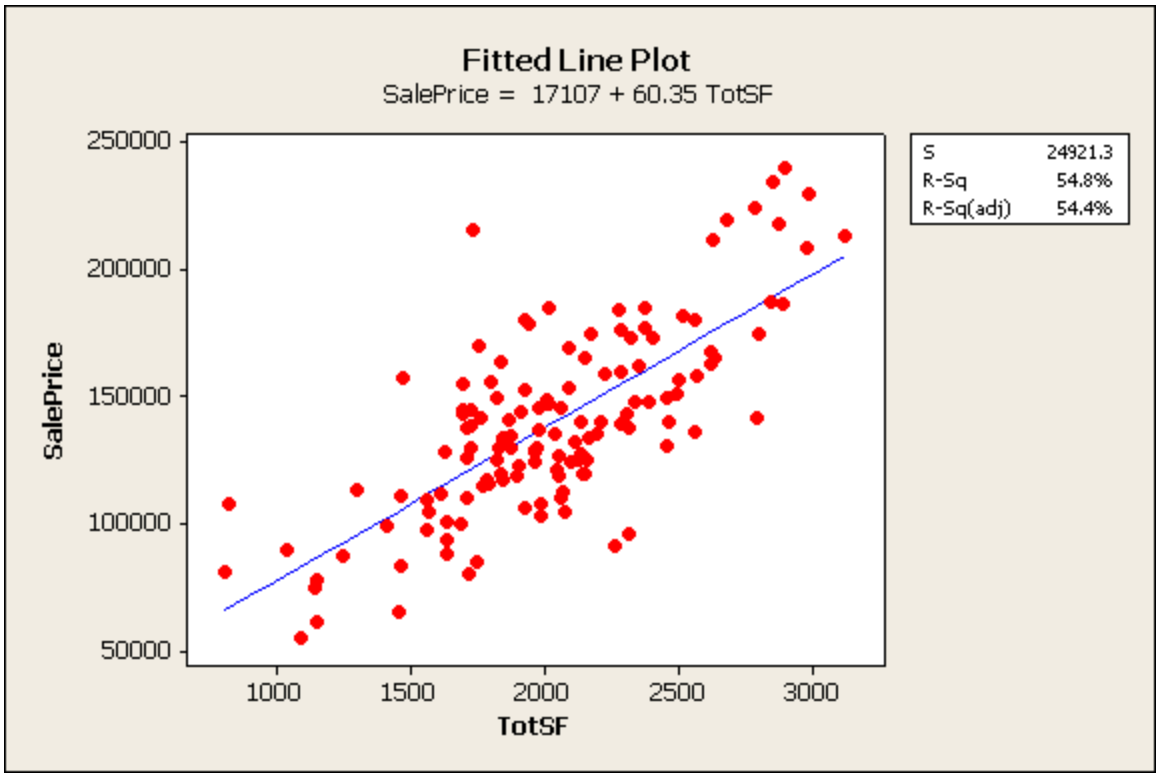

Figure 4: Fitted line plot predicting Sale Price from Total Square Footage (Ames Data)

Slightly more advanced classes may choose to discuss the topic of multiple regression. Many of the continuous variables in the data set are linearly related to the sales price of the house (one would expect the assessor's office to only collect relevant information) and there are numerous combinations which could be used together. As an example of a more advanced multiple regression model (Figure 5), I have included a mix of continuous, discrete, and nominal categorical variables. The nominal variable was created by recoding the number of fireplaces into a simple Yes/No (1/0) dummy variable. I have also included the 4-in-1 plot available on Minitab which allows for easy discussion of the assumptions of linear regression.

Regression Analysis: SalePrice versus Lot Area, Total Bsmt SF, ...

```
The regression equation is
SalePrice = 7851 + 1.72 Lot Area + 41.2 Total Bsmt SF + 40.8 Gr Liv Area
            + 20952 Garage Cars + 8379 FireYN


Predictor        Coef  SE Coef     T      P
Constant         7851    10018  0.78  0.435
Lot Area       1.7180   0.5306  3.24  0.002
Total Bsmt SF  41.188    5.795  7.11  0.000
Gr Liv Area    40.830    8.706  4.69  0.000
Garage Cars     20952     2895  7.24  0.000
FireYN           8379     3861  2.17  0.032

S = 20109.8   R-Sq = 71.4%   R-Sq(adj) = 70.3%
```

Figure 5: Minitab output for multiple regression model (Ames Data)

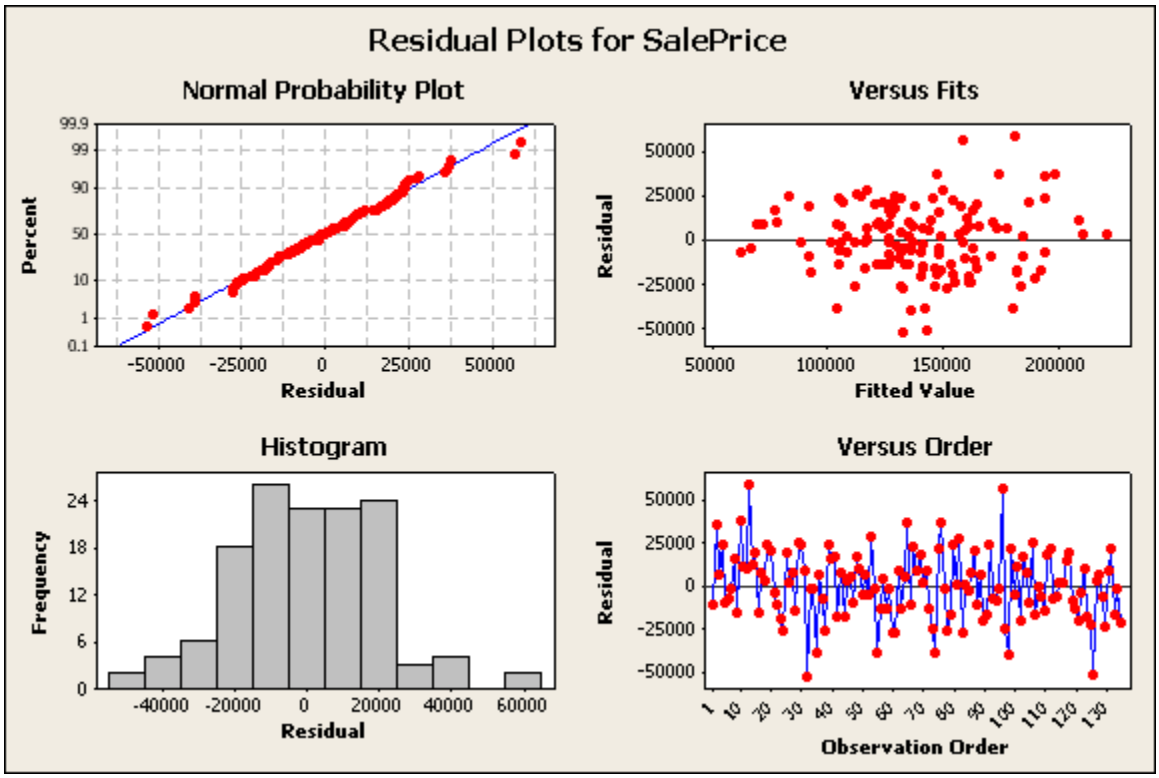

Figure 6: Minitab (4-in-1) Plot for assumption verification (Ames Data)

For instructors who cover nominal variables in their class, I would suggest incorporating the neighborhood variable into their models by converting it to a set of dummy (indicator) variables. I have found that the coefficients for the continuous variables tend to have values with more realistic interpretations when used in conjunction with the neighborhood variable.

> ***Alternative Application: As an alternative to letting students create complicated models from the large number of variables, one could challenge them to try and get the most information out of very simple models or plots. The simple plot in Figure 7 uses only the total area of a house (basement area + first and second floor area) and the type of sale but summarizes much of the variation in sale price. One can see that larger houses cost more with a bonus being paid for "Partial" sales (new homes only partially complete at last assessment) and a discount being paid for "Abnormal" sales (short sales & foreclosures).***

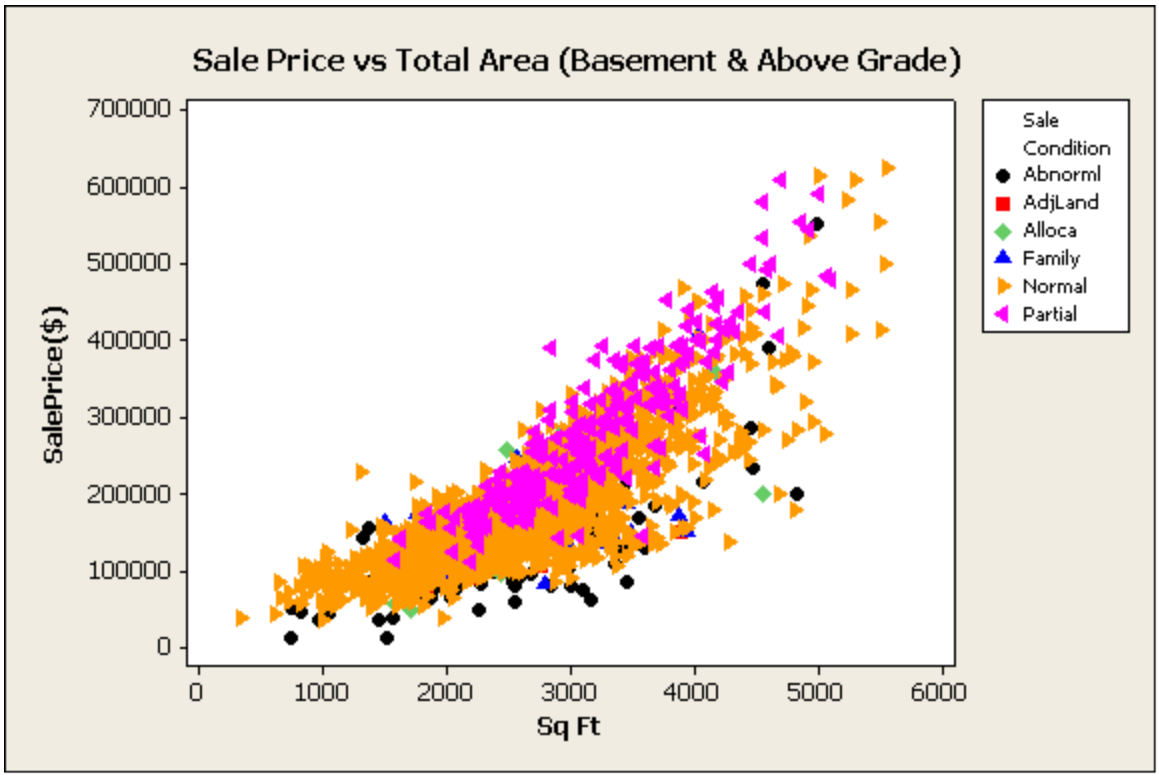

Figure 7: Scatterplot of Sale Price versus Total Home Area by Sale Condition (Ames Data)

6. Issues for Higher-Level Classes

While the variables and coding within the data set are relatively straightforward and understandable, several variables may present challenges for students who try to incorporate them into their model. As a group, the 23 ordinal variables present special difficulties. Almost all of these variables are quality related, with the expectation that higher categories should yield a coefficient at or above the previous category. In some of my initial modeling, I found that the estimated coefficients for a number of these categories did not follow this rule, likely due to interrelations with other variables within the model. While not incorrect, this situation leads to confusing interpretations (lower quality is better?) for the students. I found that some of these anomalies could be remedied by collapsing some of the larger five and ten point quality scales into fewer categories.

> ***Helpful Hint: If your students will be running a General Linear Model (rather than regression) you may want to consider revising some of the nominal and ordinal codes. Some programs (e.g. R or SAS) allow you to designate a baseline group for a variable, but other programs such as MINITAB and SPSS assume that the last category (alphabetically) will be the basis for estimation of coefficients. Manipulating the names so that the most common or lowest group is the baseline helps students in interpreting how various levels of the variable affect price. Whenever possible I tried to set variable codes so this would occur (hence my changing "Average" to "Typical/Average" allows for "T" to be the last alphabetically) but in some cases no reasonable coding could be***

found. A quick fix is to add a "z" to the beginning of the group name that you wish to be the base.

Closely related to this issue are the 14 discrete variables which are typically quantifying various types of rooms or items within a house. When treated categorically many of these items exhibit the same inconsistency as the ordinal variables (a fourth bathroom actually detracts from the value of a house?). Potentially these inconsistencies can be relieved by treating the variables as covariates which results in equal increases per item (I found this worked well with number of bathrooms) or by once again collapsing the number of categories (which worked well with number of cars per garage).

Some instructors may choose to remedy these difficulties before giving the data set to their students (using my suggested changes), but others may view this as an opportunity to discuss such anomalies with their students. Such situations can lead to rich discussion within the classroom related to collecting data, choosing variable types, and the assumptions that go with those choices.

7. Conclusion

Although I have not yet used the Ames data set in my regression class (I will in my next teaching rotation), I am confident that substituting it for the Boston data will serve my needs well. The substantial number of observations and easily understood variables offer the students many opportunities to exhibit the skills they have learned during the semester. Although I have not discussed any specific models for the large data set, I would be remiss to not admit to having spent a fair amount of time playing with the data. To give readers a benchmark, I found about 80% of the variation in residential sales price can be explained by simply taking into consideration the neighborhood and total square footage (*TOTAL BSMT SF + GR LIV AREA*) of the dwelling. On the other extreme, I have constructed a model with 36 variables (some of my own creation through recoding and interactions), all significant at the .05 level, which explains 92% of the variation in sales. While I would consider this model overly complicated, it yielded intuitively appealing coefficients where positive attributes (such as being near a park) added to the value of the home and negative attributes (such as being adjacent to a railroad) subtracted from the value. I have chosen not to share the specifics of these models here (to foil my more motivated students) but would willingly share my results with any interested instructors.

I believe the data set has unlimited potential for incorporation into a lower level statistical class. Beyond the obvious descriptive statistics, correlations and simple regression models, the diversity of variables offers opportunities to demonstrate almost all of the plots, tables, and cross-tabulations typically covered in an introductory class. My hope in sharing this data set is that instructors will find many innovative ways to incorporate it within their classes.

Acknowledgements

I would like to acknowledge Truman State University for granting me sabbatical leave for the academic year allowing me time to pursue this and other activities. I would like to thank the Ames City Assessor's Office for their willingness to share these data and for their time and effort in addressing my questions as I finalized this data set. Additionally, I would like to thank the Iowa State University StatCom student group for first introducing me to the existence of this data source.

References

Beacon (2011). *Local Government GIS for the Web*. Retrieved March 24, 2011 from http://beacon.schneidercorp.com/

City of Ames Iowa (2002). *Ames City Assessor Homepage*. Retrieved March 24, 2011 from http://www.cityofames.org/assessor/

Carnegie Mellon University (2011). *StatLib---Datasets Archive*. Retrieved April 21, 2011 from http://lib.stat.cmu.edu/datasets/boston

Dielman, T.E. (2005). *Applied Regression Analysis*. Fourth edition, Belmont, CA: Brooks/Cole.

GAISE, (2005). Guidelines for Assessment and Instruction in Statistics Education (GAISE) college report. Retrieved March 24, 2011 from The American Statistical Association (ASA): http://www.amstat.org/education/gaise/GAISECollege.htm.

Harrison, D. and Rubinfeld, D. L. (1978). "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, 5, 81-102.

Neter, J., Kutner, M., Nachtsheim, C. and Wasserman, W. (1996). *Applied Linear Statistical Models*. Fourth edition, Chicago, IL: Irwine.

Pardoe, I. (2008). "Modeling home prices using realtor data", *Journal of Statistics Education*, Volume 16, Number 2. http://www.amstat.org/publications/jse/v16n2/datasets.pardoe.html

Woodard, R. and Leone, J. (2008). "A random sample of Wake County, North Carolina residential real estate plots", *Journal of Statistics Education*, Volume 16, Number 3 (2008). http://www.amstat.org/publications/jse/v16n3/datasets.woodard.html

---

Dean De Cock
Truman State University
100 E. Normal St., Kirksville, MO, 63501
mailto:decock@truman.edu
660-785-7367 (o)
660-785-4251(f)

Volume 19 (2011) | Archive | Index | Data Archive | Resources | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Guidelines for Readers/Data Users | Home Page | Contact JSE | ASA Publications