Modeling home prices using realtor data

Iain Pardoe

Lundquist College of Business

University of Oregon

September 29, 2007

Abstract

It can be challenging when teaching regression concepts to find interesting real-life datasets that allow analyses that put all the concepts together in one large example. For example, concepts like interaction and predictor transformations are often illustrated through small-scale, unrealistic examples with just one or two predictor variables that make it difficult for students to appreciate how these concepts might be applied in more realistic multi-variable problems. This article addresses this challenge by describing a complete multiple linear regression analysis of home price data that covers many of the usual regression topics, including interaction and predictor transformations. The analysis also contains useful practical advice on model building—another topic that can be hard to illustrate realistically—and some novel statistical graphics for interpreting regression model results. The analysis was motivated by the sale of a house by the author, while the statistical ideas discussed range from those suitable for a second college statistics course to those typically found in more advanced linear regression courses.

Key words: Graphics; Indicator variables; Interaction; Linear regression; Model building; Quadratic; Transformations.

1 Introduction

This article describes a complete multiple linear regression analysis of home price data for a city in Oregon, USA in 2005. At the time the data were collected, the author was preparing to place his

house on the market and it was important to come up with a reasonable asking price. Whereas realtors use experience and local knowledge to subjectively value a house based on its characteristics (size, amenities, location, etc.) and the prices of similar houses nearby, regression analysis provides an alternative that more objectively models local house prices using these same data. Better still, realtor experience can help guide the modeling process to fine-tune a final predictive model.

The analysis described in this article contains many elements covered in typical regression components of college statistics courses such as exploratory data analysis, indicator variables for coding qualitative information, model building, hypothesis testing, diagnostics, and model interpretation. Moreover, the analysis also contains a compelling application of some more challenging topics including predictor interactions, predictor transformations, and understanding model results through the use of graphics. The article discusses statistical ideas that range from those suitable for the regression component of a second college statistics course to those typically found in more advanced multiple linear regression courses. The author has used the material in his own second statistics course (taken by business undergraduates at the University of Oregon). The example generates a lot of discussion with students able to strongly relate to questions about house values either directly or (more typically) through their parents' homes. Students can engage with this application for a variety of reasons, for example, thinking about the relative values of different house characteristics (Is an additional bathroom valued more than an additional bedroom?), or, as suggested by a referee, predicting the sale price for a house similar to the one that they grew up in.

The article, which is based on a case study in Pardoe (2006), is organized as follows: Section 2 describes the dataset; Section 3 outlines exploratory analysis of the data; Section 4 discusses development of a suitable multiple linear regression model; Section 5 provides results and conclusions from the analysis; Section 6 describes how to construct graphs to better understand the results; and Section 7 concludes with ideas for extending the analysis in class or in student assignments.

2 Data Description

The data file contains information on 76 single-family homes in Eugene, Oregon during 2005. The data were provided by Victoria Whitman, a realtor in Eugene. We wish to model sale prices of single-family homes ($Price$, in thousands of dollars) using the following predictor variables:

$Size$ = floor size (thousands of square feet)
Lot = lot size category (from 1 to 11—explained below)
$Bath$ = number of bathrooms (with half-bathrooms counting as 0.1—explained below)
Bed = number of bedrooms (between 2 and 6)
Age = age (standardized: $(\text{year built} - 1970)/10$—explained below)
$Garage$ = garage size (0, 1, 2, or 3 cars)
$Active$ = indicator for "active listing" (reference: pending or sold)
$Edison$ = indicator for Edison Elementary (reference: Edgewood Elementary)
$Harris$ = indicator for Harris Elementary (reference: Edgewood Elementary)
$Adams$ = indicator for Adams Elementary (reference: Edgewood Elementary)
$Crest$ = indicator for Crest Elementary (reference: Edgewood Elementary)
$Parker$ = indicator for Parker Elementary (reference: Edgewood Elementary)

It seems reasonable to expect that homes built on properties with a large amount of land area command higher sale prices than homes with less land, all else being equal. However, an increase in land area of (say) 2000 square feet from 4000 to 6000 should probably make a larger difference (to sale price) than going from $24{,}000$ to $26{,}000$. Thus, realtors have constructed lot size "categories," which in their experience correspond to approximately equal-sized increases in sale price. The categories (variable Lot) used in this dataset are:

Lot size	0-3k	3-5k	5-7k	7-10k	10-15k	15-20k	20k-1ac	1-3ac	3-5ac	5-10ac	10-20ac
Category	1	2	3	4	5	6	7	8	9	10	11

Lot sizes ending in "k" represent thousands of square feet, while "ac" stands for acres—there are $43{,}560$ square feet in an acre. This will prove to be important when we come to use Lot in a multiple linear regression model in Section 4. In a multiple linear regression model, predictors necessarily have "linear" impacts on the response variable ($Price$), such that a one unit change in Lot is associated with a fixed change in $Price$, whether going from categories 2 to 3 or 7 to 8. By contrast, using actual lot size in square feet in a model would produce less realistic results in which an increase in land area from 4000 to 6000 square feet would be no different (in terms of sale price) than going from $24{,}000$ to $26{,}000$.

Realtors have also recognized that "half-bathrooms" (without a shower or bath-tub) are not valued by home-buyers nearly as highly as "full" bathrooms. In fact, it appears that their value is usually not even one-half of a full bathroom and tends to be closer to one-tenth of their value—this is reflected in the definition of the variable $Bath$, which records half-bathrooms with the value 0.1.

Different housing markets value properties of various ages in different ways. This particular market has a mix of homes that were built from 1905 to 2005, with an average of around 1970. In the realtor's experience, both very old homes and very new homes tend to command a price premium relative to homes of "middle age" in this market. Thus, a quadratic effect might be expected for an age variable in a multiple linear regression model to predict price. To facilitate this we calculate a rescaled "age" variable from the "year built" variable by subtracting 1970 (the approximate mean) and dividing by 10. The resulting Age variable has a mean close to zero and a standard deviation just over 2, and leads to an intuitive interpretation for Age—it represents the number of decades away from 1970.

This dataset includes homes that have recently sold, where $Price$ represents the final sale price. However, it also includes homes that are "active listings"—homes offered for sale but which have not sold yet. At the time these data were collected, the final sale price of a home could sometimes be considerably less than the price for which it was initially offered. The dataset also includes homes that were "pending sales" for which a sale price had been agreed but paperwork still needed to be completed. To account for possible differences between final sale prices and offer prices, we define an indicator variable, $Active$, to model differences between actively listed homes ($Active = 1$) and pending or sold homes ($Active = 0$).

This particular housing market comprises a number of different neighborhoods, each with potentially different levels of housing demand. The strongest predictor of demand that is available with this dataset relates to the nearest school for each home. The housing market is contained within the geographic boundaries of a single high school, but there are six different elementary schools within this area. Thus, we define five indicator variables to serve as a proxy for the geographic neighborhood of each home. The most common elementary school in the dataset is Edgewood Elementary School so we select this to be the "reference level." The indicator variables $Edison$ to $Parker$ then represent differences of each of the schools from Edgewood.

3 Exploratory Data Analysis

Figure 1 displays a scatterplot matrix of the quantitative variables in the dataset (all of the figures and results in this article were obtained using R statistical software). While a number of bivariate

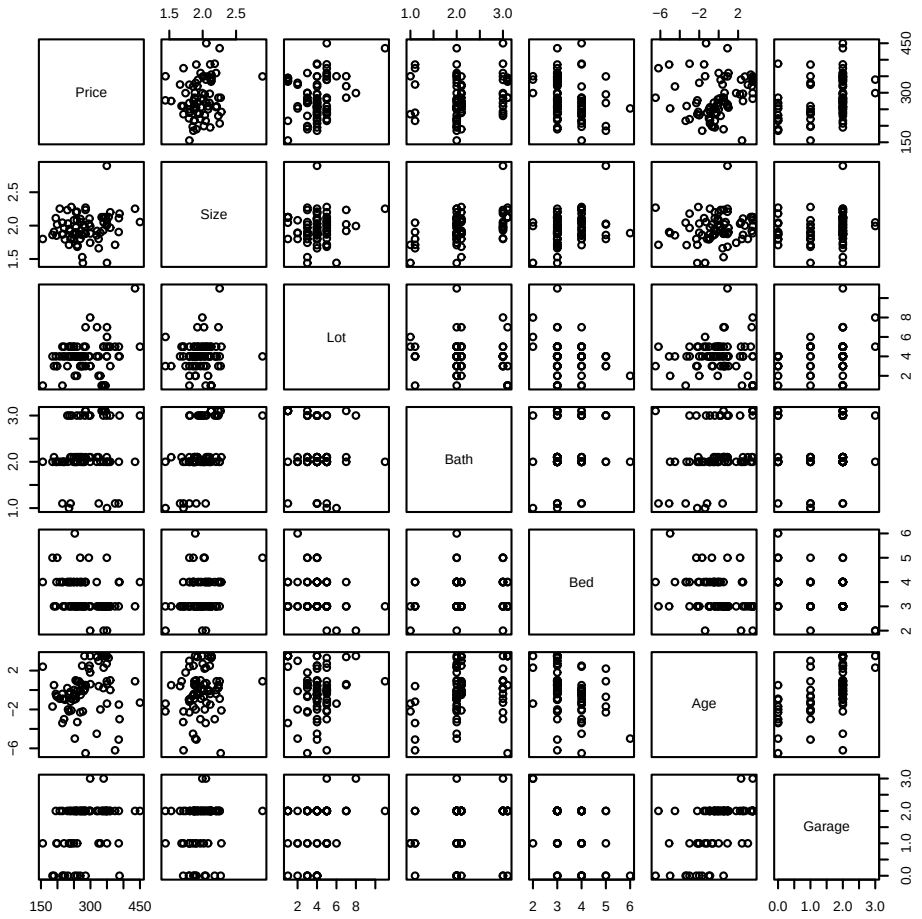

Figure 1: *Scatterplot matrix of the quantitative variables.*

relationships are evident, these tell us little about the likely form of a useful multiple linear regression model. The plots show a number of points that stick out from the dominant patterns. None of these values are so far from the remaining values that they are likely to cause a problem with subsequent analysis, but it is worth making a note of them just in case. Home 76 is much larger than the rest, while home 74 has a larger lot size than the rest (and is quite expensive). Home 35 is the only one with six bedrooms, while home 54 is the oldest home. Homes 21 and 47 are the only ones with three-car garages, while home 2 is the most expensive and home 5 is the cheapest.

Figure 2 displays boxplots of $Price$ versus the two qualitative variables. Home prices are less variable and have a higher median for active listings relative to recently sold homes (or pending sales). Prices tend to be higher in neighborhoods near to Edison and Harris schools, and lower for

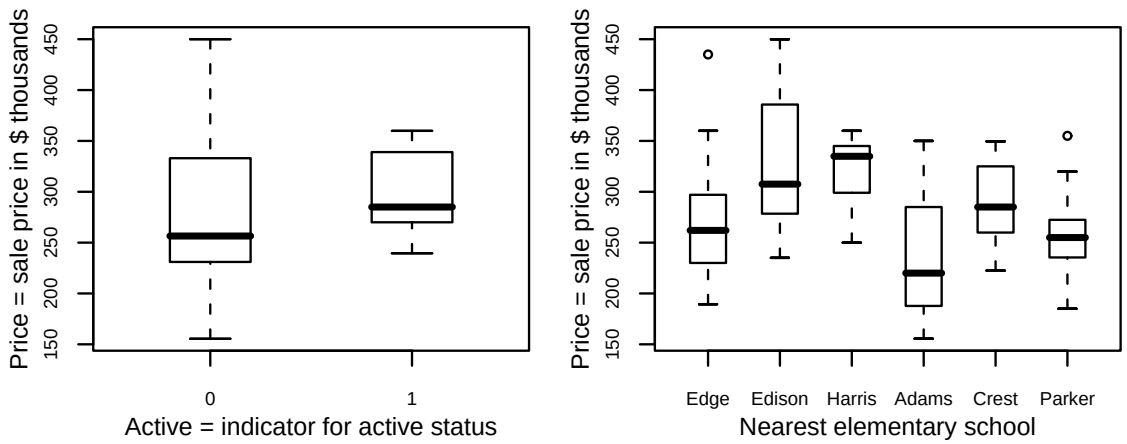

Figure 2: *Boxplots of $Price$ versus $Active$ (left) and $Price$ versus elementary school (right).*

Adams, with the other three schools broadly similar in a "moderate" range. Home 74 is particularly expensive for a typical Edgewood home, while home 40 is similarly expensive for a typical Parker home. The numbers of homes in neighborhoods near to Crest (6 homes) and Adams (3 homes) are relatively small, which may limit our ability to say much about systematic differences in house prices for these two neighborhoods. We return to this question at the end of Section 4.

Keep in mind that these observations about qualitative predictors do not take into account the quantitative predictors, $Size$, Lot, and so on.

4 Regression Model Building

Having got a feel for the data, we next want to apply multiple linear regression modeling to see whether the various house characteristics allow us to model sale prices with any degree of accuracy. All of the topics considered in this section (and also Section 5)—namely R^2, adjusted R^2, regression standard error, residual analysis, indicator variables, variable transformations, interactions, regression assumptions, variable selection, and nested model F-tests—would typically be covered during the regression component of a second college statistics course.

We first try a model with each of the predictors "as is" (no transformations or interactions):

$$\begin{aligned} \mathrm{E}(Y) &= b_0 + b_1 Size + b_2 Lot + b_3 Bath + b_4 Bed + b_5 Age + b_6 Garage \\ &\quad + b_7 Active + b_8 Edison + b_9 Harris + b_{10} Adams + b_{11} Crest + b_{12} Parker. \end{aligned}$$

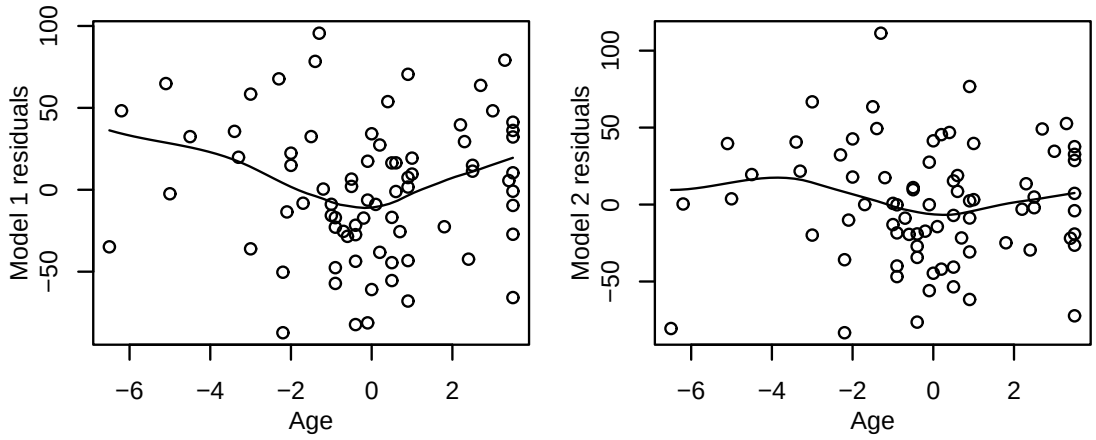

Figure 3: *Residual plots for the first model (left) and second model (right), both with* Age *on the horizontal axis and loess fitted lines superimposed.*

This model results in values of $R^2 = 0.530$ (coefficient of determination) and $s = 45.1$ (regression standard error) but isn't very satisfactory for a number of reasons. For example, the residuals from this model fail to satisfy the zero mean (linearity) assumption in a plot of the residuals versus Age, displaying a relatively pronounced curved pattern. The left-hand plot in Figure 3 displays this plot, together with a *loess fitted line* that provides a graphical representation of the average value of the residuals as we move across the plot (i.e., as Age increases). Pardoe (2006, p. 107) and Cook and Weisberg (1999, p. 44) provide more details on the use of loess fitted lines for assessing patterns in scatterplots—this might be a more suitable topic for a more advanced regression course.

To attempt to correct this failing, we will add an Age^2 transformation to the model, which as discussed above was also suggested from the realtor's experience. The finding that the residual plot with Age has a curved pattern does not necessarily mean that an Age^2 transformation will correct this problem, but it is certainly worth trying.

In addition, both $Bath$ and Bed have relatively large individual t-test p-values in this first model, which appears to contradict the notion that home prices should increase with the number of bedrooms and bathrooms. However, the relationship with bedrooms and bathrooms may be complicated by a possible interaction effect. For example, adding extra bathrooms to homes with just two or three bedrooms might just be considered a waste of space and so have a negative impact on price. Conversely, there is a clearer benefit for homes with four or five bedrooms to have more than one bathroom and so adding bathrooms for these homes probably has a positive impact on

price. To model such a relationship we will add a $Bath \times Bed = BathBed$ interaction term to the model. Therefore, we next try the following model:

$$\begin{aligned} \mathrm{E}(Y) &= b_0 + b_1 Size + b_2 Lot + b_3 Bath + b_4 Bed + b_{34} BathBed \\ &\quad + b_5 Age + b_{52} Age^2 + b_6 Garage + b_7 Active + b_8 Edison \\ &\quad + b_9 Harris + b_{10} Adams + b_{11} Crest + b_{12} Parker. \end{aligned}$$

This model results in values of $R^2 = 0.599$ and $s = 42.4$ and has residuals that appear to satisfy the four regression model assumptions of zero mean (linearity), constant variance, normality, and independence reasonably well (the residual plot with Age on the horizontal axis is displayed as the right-hand plot in Figure 3 but the other residual plots are not shown). However, the model includes some terms with large individual t-test p-values, suggesting that perhaps it is more complicated than it needs to be and includes some redundant terms. In particular, the last three elementary school indicators ($Adams$, $Crest$, and $Parker$) have p-values of 0.310, 0.683, and 0.389. We conduct a nested model F-test (also known as an "analysis of variance" test or "extra sum of squares" test) to see whether we can safely remove these three indicators from the model without significantly worsening its fit (at a 5% significance level):

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error	Change Statistics F-stat	df1	df2	Pr(>F)
3	0.767 [a]	0.588	0.518	41.907				
2	0.774 [b]	0.599	0.507	42.370	0.537	3	61	0.659

[a] Predictors: (Intercept), $Size$, Lot, $Bath$, Bed, $BathBed$, Age, Age^2, $Garage$, $Active$, $Edison$, $Harris$.

[b] Predictors: (Intercept), $Size$, Lot, $Bath$, Bed, $BathBed$, Age, Age^2, $Garage$, $Active$, $Edison$, $Harris$, $Adams$, $Crest$, $Parker$.

Since the p-value of 0.659 is more than any sensible significance level, we cannot reject the null hypothesis that the last three school indicator regression parameters are all zero. In addition, removing these three indicators improves the values of adjusted R^2 (from 0.507 to 0.518) and s (from 42.4 to 41.9). Removing these three indicators from the model means that the school reference level now comprises Edgewood, Adams, Crest, and Parker (so that there are no systematic differences between these four schools with respect to home prices).

5 Results and Conclusions

Thus, a final model for these data is

$$\begin{aligned} \mathrm{E}(Y) &= b_0 + b_1 Size + b_2 Lot + b_3 Bath + b_4 Bed + b_{34} BathBed \\ &\quad + b_5 Age + b_{52} Age^2 + b_6 Garage + b_7 Active + b_8 Edison + b_9 Harris. \end{aligned}$$

Statistical software output for this model is:

Model Summary

Model	Multiple R	R Squared	Adjusted R Squared	Regression Std. Error
3	0.767 [a]	0.588	0.518	41.907

[a] Predictors: (Intercept), $Size$, Lot, $Bath$, Bed, $BathBed$, Age, Age^2, $Garage$, $Active$, $Edison$, $Harris$.

Parameters [a]

Model		Estimate	Std. Error	t-stat	Pr(>|t|)
3	(Intercept)	332.478	106.599	3.119	0.003
	$Size$	56.719	27.974	2.028	0.047
	Lot	9.917	3.438	2.885	0.005
	$Bath$	−98.156	42.666	−2.301	0.025
	Bed	−78.910	27.752	−2.843	0.006
	$BathBed$	30.390	11.878	2.559	0.013
	Age	3.301	3.169	1.042	0.302
	Age^2	1.641	0.733	2.238	0.029
	$Garage$	13.119	8.285	1.583	0.118
	$Active$	27.424	10.988	2.496	0.015
	$Edison$	67.062	16.822	3.987	0.000
	$Harris$	47.273	14.844	3.185	0.002

[a] Response variable: $Price$.

The estimated regression equation is therefore:

$$\begin{aligned} \widehat{Price} &= 332.48 + 56.72 Size + 9.92 Lot - 98.16 Bath - 78.91 Bed \\ &\quad + 30.39 BathBed + 3.30 Age + 1.64 Age^2 + 13.12 Garage \\ &\quad + 27.42 Active + 67.06 Edison + 47.27 Harris. \end{aligned}$$

This final model results in residuals that appear to satisfy the four regression model assumptions of zero mean (linearity), constant variance, normality, and independence reasonably well (residual plots not shown). Also, each of the individual t-test p-values is below the usual 0.05 threshold (including $Bath$, Bed, and the $BathBed$ interaction), except Age (which is included to retain hierarchy since Age^2 is included in the model) and $Garage$ (which is nonetheless retained since its 0.118 p-value is low enough to suggest a potentially important effect).

The model can explain 58.8% of the variation in price, and predictions using the model are likely to be accurate to within approximately $\pm\$83{,}800$ (at a 95% confidence level). To put this in context, prices in this dataset range from $\$155{,}000$ to $\$450{,}000$. This still leaves more than 40% of the variation in price unexplained by the model, which suggests that the dataset predictors can only go so far in helping to explain and predict home prices in this particular housing market. Variables not measured that could account for the remaining 41.2% of the price variation might include other factors related to the geographical neighborhood, condition of the property, landscaping, and features such as updated kitchens and fireplaces.

A potential use for the model might be to narrow the range of possible values for the asking price of a home about to be put on the market. For example, consider a home with the following features: 1879 square feet, lot size category 4, two and a half bathrooms, three bedrooms, built in 1975, two-car garage, and near to Parker Elementary School (this was the author's house at the time). A 95% prediction interval ignoring the model comes to ($\$164{,}800$, $\$406{,}800$); this is based on the formula: sample mean $\pm$ t-percentile $\times$ sample standard deviation $\times\sqrt{(1+1/n)}$. By contrast, a 95% prediction interval using the model results comes to ($\$197{,}100$, $\$369{,}000$), which is about 70% the width of the interval ignoring the model. A realtor could advise the vendors to price their home somewhere within this range depending on other factors not included in the model (e.g., toward the upper end of this range if the home is on a nice street, the property is in good condition, and some landscaping has been done to the yard). As is often the case, the regression analysis results are more effective when applied in the context of expert opinion and experience.

These results also illustrate that "prediction is hard;" whereas the final model might be considered quite successful in terms of its ability to usefully explain nearly 60% of the variation in sale prices, the 95% prediction interval of ($\$197{,}100$, $\$369{,}000$) is perhaps disappointingly wide. Students might like to consider ways to tighten up this interval, for example, by collecting more

data observations or thinking of new predictor variables.

A further use for the model might be to utilize the specific findings relating to the effects of each of the predictors on the price. Since $\hat{b}_1 = 56.72$, we expect sale price to increase by $\$5672$ for each 100 square foot increase in floor size, all else held constant. Similarly, since $\hat{b}_2 = 9.92$, we expect sale price to increase by $\$9920$ for each one-category increase in lot size, all else held constant. Similarly, since $\hat{b}_6 = 13.12$, we expect sale price to increase by $\$1312$ for each vehicle increase in garage size, all else held constant.

The preceding discussion contains most of the standard topics that would typically be covered during the regression component of a second college statistics course. Such courses sometimes also cover more "advanced" topics, such as the role that individual data observations can play in a multiple linear regression model (e.g., outliers or high leverages); these topics would certainly be covered in a more advanced course dealing only with regression. To illustrate, calculation of studentized residuals, leverages, and Cook's distances can help to identify overly influential observations. Finding such observations can suggest the need to investigate possible data errors, to add additional predictors to the model, to respecify the model in some other way, or to consider removing the influential observations from the dataset. However, in this case none of the final model studentized residuals are outside the ± 3 range, and so none of the observations would probably be considered outliers. Home 76 (with a large floor size) has the highest leverage, although home 54 (the oldest home) is not far behind. These two homes also have the two highest Cook's distances, although neither is above a 0.5 threshold (see Cook and Weisberg 1999, p. 358), and neither dramatically changes the regression results if excluded.

6 Predictor Effect Plots

Interpretation of the parameter estimates for $Bath$, Bed, and Age are complicated somewhat by their interactions and transformations. In such circumstances it can be helpful to use statistical graphics to help understand the model results. This section describes "predictor effect plots," line plots that show graphically how a regression response variable (home price in this case) is associated with changes to the predictor variables. We will see that for the variables $Size$, Lot, and $Garage$, these line plots simply represent the corresponding parameter estimates as straightforward

slopes. However, in the case of $Bath$, Bed, and Age the plots provide additional insights due to the presence of complicating interaction effects and transformations.

Note that estimated regression parameters cannot usually be given *causal* interpretations. The regression modeling described in this article can really only be used to quantify relationships and to identify whether a change in one variable is associated with a change in another variable, not to establish whether changing one variable "causes" another to change. The term "predictor effect" in this section indicates how a regression model expects $Price$ to change as each predictor changes (and all other predictors are held constant), but without suggesting at all that this is some kind of causal effect.

The basic ideas behind predictor effect plots are sufficiently straightforward that they could be comfortably covered in the regression component of a second college statistics course. First, consider how $Price$ changes as $Size$ changes. Since $Size$ is not included in any interaction terms, we can isolate this change in $Price$ when we hold the remaining predictors constant (say, at sample mean values for the quantitative predictors and zero for the indicator variables). Then the "$Size$ effect on $Price$" is

$$Size \text{ effect on } Price = 135.1 + 56.72Size.$$

The value 56.72 comes directly from the $Size$ part of the estimated regression equation, while the value 135.1 results from plugging in the sample means for Lot, $Bath$, Bed, Age, and $Garage$, and zero for $Active$, $Edison$, and $Harris$ to the rest of the equation. This $Size$ effect then represents how $Price$ changes as $Size$ changes for homes with average values for $Lot, \ldots, Garage$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods.

We can then construct a line plot with this $Size$ effect on the vertical axis and $Size$ on the horizontal axis—the left-hand plot in Figure 4 illustrates. Over the range of values in the dataset, floor size increases from approximately 1440 to 2900 square feet are associated with price increases from approximately $\$215$k to $\$300$k on average (for homes with average values for $Lot, \ldots, Garage$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods). Homes with other values for $Lot, \ldots, Garage$, or that are in other neighborhoods, tend to have price differences of a similar magnitude for similar changes in floor size (although the sales prices of individual homes will depend on those values of $Lot, \ldots, Garage$ and neighborhood)—the predictor effect plot would simply have different values on the vertical axis, but the slope of the line would be the same.

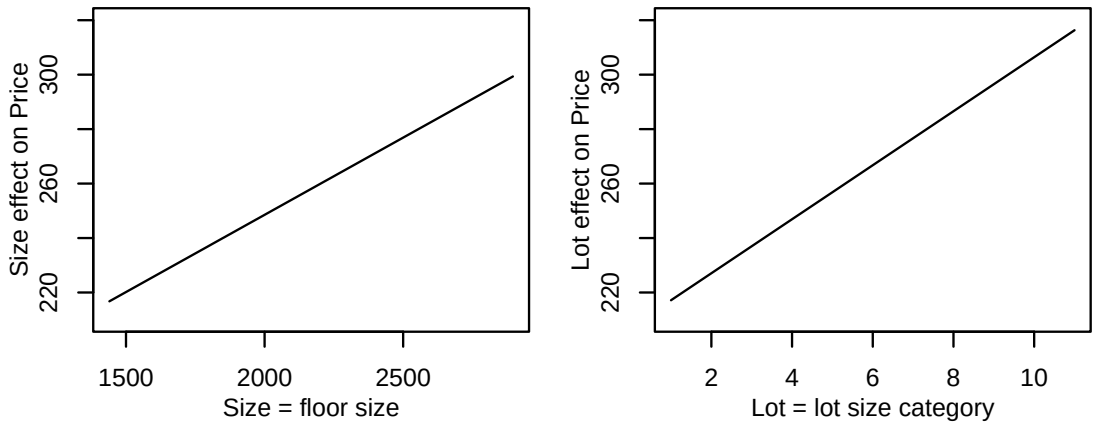

Figure 4: *Predictor effect plots for $Size$ and Lot in the home prices example. In the left plot, the $Size$ effect on $Price$ of $135.1 + 56.72Size$ is on the vertical axis, while $Size$ is on the horizontal axis. In the right plot, the Lot effect on $Price$ of $207.2 + 9.92Lot$ is on the vertical axis, while Lot is on the horizontal axis.*

Similarly, the "Lot effect on $Price$" is

$$Lot \text{ effect on } Price = 207.2 + 9.92Lot,$$

which is illustrated in the right-hand plot in Figure 4. Lot size category increases from 1 to 11 are associated with price increases from approximately $\$215$k to $\$315$k on average (for homes with average values for $Size$, $Bath$, Bed, Age, and $Garage$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods). Again, homes with other predictor values tend to have similar magnitude price differences for similar changes in lot size.

The "$BathBed$ effect on $Price$" involves an interaction:

$$BathBed \text{ effect on } Price = 504.2 - 98.16Bath - 78.91Bed + 30.39BathBed.$$

The left-hand plot in Figure 5 shows a line plot with this $BathBed$ effect on $Price$ on the vertical axis, $Bath$ on the horizontal axis, and lines marked by the value of Bed. In homes with just two or three bedrooms, additional bathrooms are associated with lower prices (holding all else constant), particularly two-bedroom homes. Conversely, in homes with four or five bedrooms, additional bathrooms are associated with higher prices (all else constant), particularly five-bedroom homes. The scale on the plot shows the approximate magnitude of average prices for different numbers

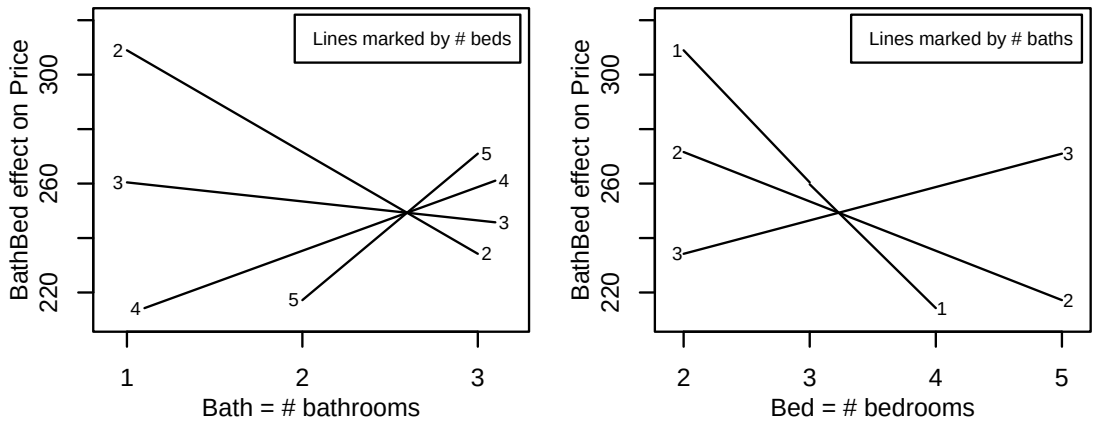

Figure 5: *Predictor effect plots for $Bath$ and Bed in the home prices example. In the left plot, the $BathBed$ effect on $Price$ of $504.2 - 98.16Bath - 78.91Bed + 30.39BathBed$ is on the vertical axis while $Bath$ is on the horizontal axis and the lines are marked by the value of Bed. In the right plot, the $BathBed$ effect on $Price$ is on the vertical axis while Bed is on the horizontal axis and the lines are marked by the value of $Bath$.*

of bathrooms and bedrooms (for homes with average values for $Size$, Lot, Age, and $Garage$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods). Again, homes with other predictor values tend to have price differences of a similar magnitude for similar changes in the numbers of bathrooms and bedrooms.

The right-hand plot in Figure 5 shows a line plot with the $BathBed$ effect on $Price$ on the vertical axis and Bed on the horizontal axis, and lines marked by the value of $Bath$. In homes with one or two bathrooms, additional bedrooms are associated with lower prices (all else constant), particularly one-bath homes. Conversely, in homes with three bathrooms, additional bedrooms are associated with higher prices (all else constant). The scale on the plot shows the approximate magnitudes of average prices for different numbers of bathrooms and bedrooms (for homes with average values for $Size$, Lot, Age, and $Garage$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods). Again, homes with other predictor values tend to have price differences of a similar magnitude for similar changes in the numbers of bathrooms and bedrooms.

The "Age effect on $Price$" is

$$Age \text{ effect on } Price = 246.9 + 3.30Age + 1.64Age^2.$$

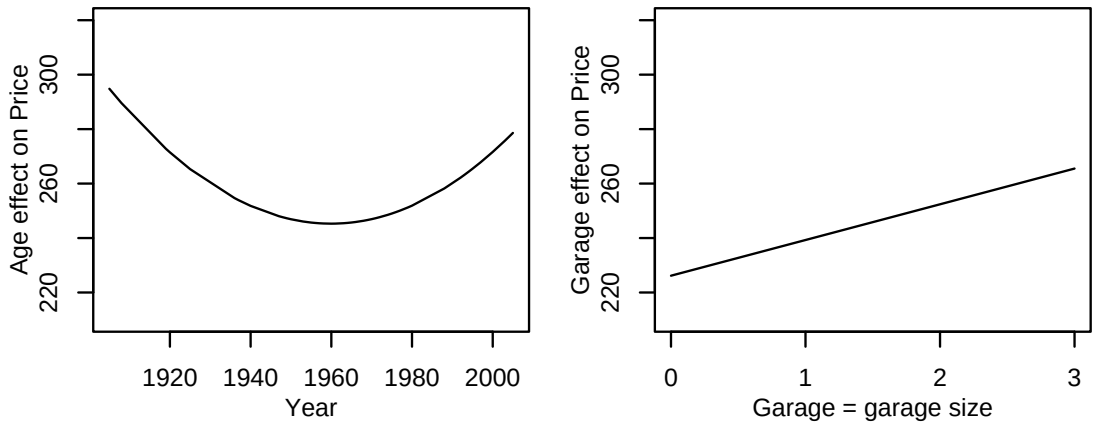

Figure 6: *Predictor effect plots for Age and $Garage$ in the home prices example. In the left plot, the Age effect on $Price$ of $246.9 + 3.30Age + 1.64Age^2$ is on the vertical axis while Age is on the horizontal axis. In the right plot, the $Garage$ effect on $Price$ of $226.2 + 13.12Garage$ is on the vertical axis while $Garage$ is on the horizontal axis.*

The left-hand plot in Figure 6 shows a line plot with this Age effect on $Price$ on the vertical axis and year on the horizontal axis. Over the range of values in the dataset, average prices decrease from a high of approximately $\$295$k to a low of approximately $\$245$k from the early 1900s to 1960, and then increase again up to approximately $\$280$k in 2005 (for homes with average values for $Size$, Lot, $Bath$, Bed, and $Garage$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods). Again, homes with other predictor values tend to have price differences of a similar magnitude for similar changes in age.

The "$Garage$ effect on $Price$" is

$$Garage \text{ effect on } Price = 226.2 + 13.12Garage,$$

which is illustrated in the right-hand plot in Figure 6. Over the range of values in the dataset, garage size increases from 0 to 3 are associated with price increases from approximately $\$225$k to $\$265$k on average (for homes with average values for $Size, \ldots, Age$ that are in the Edgewood, Adams, Crest, or Parker neighborhoods). Again, homes with other predictor values tend to have similar magnitude price differences for similar changes in garage size.

The indicator variable effects are more easily described in words. An active listing (with all else held constant) tends to be associated with a price increase of $\$27{,}400$ (perhaps suggesting

that homes offered for sale were substantially overpriced in this housing market). Finally, two elementary schools seem to offer a price premium: Edison (approximately $\$67{,}100$) and Harris (approximately $\$47{,}300$). This does not necessarily mean that it is just the proximity of these schools to a home that is associated with increased sale prices. It is more likely in this case that there are a number of features associated with these neighborhoods that tend to increase home prices (e.g., in this case Edison and Harris are both close to the major university in the state, the University of Oregon).

7 Discussion

This article has described a complete analysis—from exploratory data analysis through multiple linear regression model building—of a compelling real-life dataset on home prices. The analysis has covered many of the usual topics in a regression course, but, as is usually the case, there are additional possibilities for investigating this dataset further. The following offers some ideas for students to continue working with this dataset. Some, such as the first idea below would be reasonable for the regression component of a second college statistics course, while others, such as the final idea below, might be better suited for a more advanced course dealing only with regression.

1. It is possible that the final model could be improved by considering interactions between the quantitative predictors and the indicator variables, for example, $ActiveSize$. Investigate whether there are any such interactions that significantly improve the model.

2. Investigate whether an alternative measure of lot size might be more appropriate than the categories used in the dataset. For example, define a new predictor variable that is the natural logarithm of the mid-point of the lot size range (in thousands of square feet) represented by each category (i.e., $\ln(1.5) = 0.41$ for category 1, $\ln(4) = 1.39$ for category 2, and so on). Reanalyze the data with this new predictor in place of Lot. Do model results change drastically when you do this?

3. Investigate whether counting half-bathrooms as 0.1 is reasonable. For example, change values ending in $.1$ in the dataset to end in $.5$ instead, and reanalyze the data. Do model results change drastically when you do this?

4. Investigate whether there appear to be any systematic differences between pending sale prices and actual sales prices (all else equal). The analysis just described assumes no difference since the only indicator variable for "status" is $Active$, which is 1 for active houses and 0 for both pending sales and sold houses. Add an indicator variable that is 1 for pending sales and 0 for both active and sold houses, and reanalyze the data. Do model results change drastically when you do this?

5. The values for $Price$ are slightly skewed in a positive direction, suggesting perhaps that transforming $Price$ to $\ln(Price)$ might result in an improved multiple linear regression model. Reanalyze the data, but use $\ln(Price)$ as the response variable instead of $Price$. Interpret results, remembering that regression parameter estimates such as $\hat{b}_1$ will need to be transformed to $\exp(\hat{b}_1)-1$, where they now represent the expected proportional change in $Price$ from increasing $Size$ by one unit (all else constant). Justification for this transformation comes from the following. Partition the model predictors into X_1 and $\mathbf{X}$ (a vector representing all remaining predictors), and note that the expected (absolute) change in $Price$ after increasing X_1 by one unit (all else constant) is

$$\exp(\hat{b}_1(X_1+1) + \hat{\mathbf{b}}\mathbf{X}) - \exp(\hat{b}_1 X_1 + \hat{\mathbf{b}}\mathbf{X}) = (\exp(\hat{b}_1)-1)\exp(\hat{b}_1 X_1 + \hat{\mathbf{b}}\mathbf{X}),$$

   where $\hat{\mathbf{b}}$ is a vector representing the regression parameters for the predictors in $\mathbf{X}$. Thus, the expected proportional change in $Price$ after increasing X_1 by one unit (all else constant) is simply $\exp(\hat{b}_1)-1$.

6. Obtain similar data for a housing market near you (e.g., home listings are commonly available on the internet), and perform a regression analysis to explain and predict home prices in that market. Compare and contrast your results with the results presented here.

References

Cook, R. D. and S. Weisberg (1999). *Applied Regression Including Computing and Graphics*. Hoboken, NJ: Wiley.

Pardoe, I. (2006). *Applied Regression Modeling: A Business Approach*. Hoboken, NJ: Wiley.