

From a Theoretical Formula to a Defensible and Robust Model

The Appropriate Development and Application of Regression Models in the Context of Real Estate Appraisal

by Erin Kiella, PhD, Jennifer Pitts, MAI, and Christopher Yost-Bremm, PhD

Abstract

Hedonic regression analysis has proven to be a useful tool in numerous fields, including real estate valuation. However, while regression analysis can be useful for estimating economic real estate relationships, the method cannot be implemented in every context. Real estate especially is a nuanced asset, and special consideration must be given when constructing and interpreting a model of real estate markets. This article presents the framework for structurally modeling real estate markets and discusses factors affecting the reliability and applicability of a regression model in real estate appraisal. This article serves as a concise reference for appraisers employing a regression model, as well as for economists who are not as familiar with real estate markets, and even users of an appraisal report or economic study employing regression analysis for modeling real estate markets (i.e., a property owner, an attorney, or a judge). A practical and real-life case study illustrates how structural regression models can be used to answer valuation-related questions. Proper understanding and appropriate application of regression modeling is imperative to the credible use of the tool in real estate valuation.

Introduction: An Overview of Hedonic Regression

Hedonic regression analysis, a form of empirical analysis, uses data to test a theory or estimate a relationship.¹ *The Dictionary of Real Estate Appraisal* describes regression analysis as “a statistical method that examines the relationship between one or more independent variables and a dependent variable.”² In real estate valuation, hedonic regression analysis is a variation of the traditional sales comparison approach. It explains the sale prices of a group of properties using data on the value-influencing characteristics of each property.

Conceptually, the following equation represents a multiple regression model of sale prices (with each sale transaction in the model denoted by i):

$$\begin{aligned} \text{Sale Price}_i &= \text{Property Characteristics}_i \\ &+ \text{Sale and Market Conditions}_i \\ &+ \text{Other External Factors}_i \end{aligned}$$

In this equation, *Sale Price* is the “dependent” variable. The appraiser aims to explain property *Sale Price* variations using data on the characteristics of each property. These characteristics, termed “independent” or “explanatory” variables, are factors the appraiser identifies as contributing to the sale prices of properties in a

1. Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach* (South-Western College Publishing, 2003).
2. *The Dictionary of Real Estate Appraisal*, 7th ed. (Appraisal Institute, 2022), s.v. “regression analysis.”

particular market. They include physical property characteristics (e.g., the interior square footage, lot size, age, condition), sale and market characteristics (e.g., the year in which the sale occurred, sale conditions such as a distressed sale or nonconventional financing), and other external or locational factors (e.g., proximity to a highway, school district). The hedonic regression approach simultaneously correlates the various explanatory variables with the sale price of the property and estimates coefficients for each of the independent variables. The coefficient estimate is the contributory value of each explanatory variable to the overall sale price, similar in concept to the adjustments made in a traditional sales comparison approach.³ By combining these property characteristics into a single equation, as opposed to considering pairwise correlations (i.e., the correlation between sale price and one variable, in isolation), hedonic regression can, within limits, distinguish the economic effects of multiple individual amenities (more square footage, an additional bathroom, etc.), as well as potential disamenities (e.g., a less desirable school district, or proximity to industrial emissions).⁴

Regression in Real Estate Valuation: Predictive Versus Structural Models

Real estate appraisers typically use regression models in one of two contexts: as a predictive model or as a structural model. A predictive model uses the coefficients estimated by the model to predict the overall value of individual properties.

Predictive regression models are often used in automated valuation models (AVMs) or for tax assessment purposes.⁵

By contrast, structural models estimate the relative effects of a certain explanatory variable (e.g., size, age, the presence of a pool, garage size, proximity to an amenity or disamenity) on the dependent variable, while holding the other variables' effects constant. This type of analysis assesses whether and how the variable of interest affects the dependent variable (i.e., sale price). The model estimates (1) a coefficient representing the effect—either positive or negative as well as magnitude—of the explanatory variables on the dependent variable and (2) the statistical significance of each estimated coefficient. Regression modeling in this context has successfully been used to estimate quantitative adjustments for the sales comparison approach, to estimate the effect of an amenity or disamenity on property values (e.g., estimates of diminution in value due to environmental contamination),⁶ and in a variety of other applications. The remainder of this article discusses the use of regression for structural modeling, although the statistical issues discussed are also relevant to predictive models.

Applicability of Structural Regression Models

While hedonic regression modeling can be a useful tool in the field of real estate valuation, it is imperative that appraisers recognize the limitations of regression modeling and not assume applicability in *any* case or circumstance. The National Research Council of the National Academies warns, “The frequency with which regres-

3. See Mark R. Linné, M. Steven Kane, and George Dell, *A Guide to Appraisal Valuation Modeling* (Appraisal Institute, 2000), 49.

4. Hedonic regression is typically estimated using a statistical procedure called *Ordinary Least Squares*. More technical detail behind this procedure is discussed later in this article.

5. For more detail on predictive models, see Advisory Opinion 18, “Automated Valuation Models,” published with the *Uniform Standards of Professional Appraisal Practice* (USPAP).

6. When estimating diminution in value due to a disamenity, such as environmental contamination, the regression model estimates a single, average amount of diminution for the group of properties under study. The model is not capable of precisely estimating diminution in value specific to an individual property, nor does it vary based on the specific characteristics of a property (e.g., differences in physical and locational characteristics, extent of contamination, remediation status). Given that the result of the regression model is an average, if it is applied uniformly across properties in the area under study, diminution in value may be underestimated for some properties and overestimated for others. This problem is exacerbated when the model attempts to analyze a large, heterogeneous group of properties. An individual analysis of the physical, locational, and environmental characteristics of each property is necessary to appropriately apply any results of a regression analysis. For further discussion of these issues, see T. O. Jackson, “Real Property Valuation Issues in Environmental Class Actions,” *The Appraisal Journal* (Spring 2010): 141–149.

sion models are used is no guarantee that they are the best choice for any particular problem.”⁷ A model’s ability to yield meaningful output relies on appropriate application, proper and rigorous development, and correct interpretation of results. *The Appraisal of Real Estate* states that “credible regression modeling includes an assessment of data sufficiency, a residual analysis, an assessment of which variables should be included in a model, and model validation.”⁸

Like any tool, regression analysis can be developed poorly and used or interpreted incorrectly. Proper empirical safeguards must be instituted before, during, and after the analysis to ensure credible and reliable output. As the National Research Council admonishes, “Failure to develop the proper theory, failure to choose the appropriate variables, or failure to choose the correct form of the model can substantially bias the statistical results—that is, create a systematic tendency for an estimate of a model parameter to be too high or too low.”⁹ A user of regression methodology can efficiently quantify the contributory effects of different value-influencing characteristics of real estate. However, if estimated without forethought and proper design (termed “model specification”), such modeling can result in misleading value conclusions.

Unlike the market for other goods and services, real estate is a unique asset and subject to individualized value influences. *The Appraisal of Real Estate* describes the unique aspects of the real estate market, stating:

The efficiency of a market is tied to the behavior of buyers and sellers. ... Real estate markets can differ significantly from the markets for other goods and services and have never been considered truly efficient markets. ... Real estate products are heterogeneous, and information about real estate is often incomplete.¹⁰

As such, developing structural models to analyze economic impact on property values requires an understanding of these issues and nuances and how they may affect a regression model as well as an understanding of the tools and techniques to address them.

The Appraisal of Real Estate elaborates on the misuse of regression analysis in real estate valuation, stating, for example, that “[m]odel specification issues fall into two broad categories for valuation purposes: (1) the functional form of the relationship between the dependent variable and the independent variables and (2) the choice of variables to include in the model.”¹¹ Similarly, in his seminal textbook, *Introductory Econometrics: A Modern Approach*, Jeffrey M. Wooldridge writes, “A multiple regression model suffers from functional form misspecification when it does not properly account for the relationship between the dependent and the observed explanatory variables. Functional form misspecification leads to biased estimators.”¹²

The following sections examine in detail how these and other types of model misspecification may occur and their effect on the reliability of a model’s results. A model plagued with even one of these statistical issues can lead to unreliable conclusions. A case study is presented along with the discussion of model issues to illustrate how these potential statistical issues can be avoided or corrected in a practical application of structural regression models in a real estate valuation context.

Errors and Omissions in Hedonic Regression

The appropriateness of a hedonic regression model as a tool for an appraisal study extends beyond merely collecting data and inputting it into statistical software. The following concepts

7. Federal Judicial Center and National Research Council. *Reference Manual on Scientific Evidence*, 3rd ed. (National Academies Press, 2011), 272.

8. *The Appraisal of Real Estate*, 15th ed. (Appraisal Institute, 2020), Appendix B: 17.

9. *Reference Manual on Scientific Evidence*, 3rd ed., 312.

10. *The Appraisal of Real Estate*, 15th ed., 114.

11. *The Appraisal of Real Estate*, 15th ed., Appendix B: 9.

12. Wooldridge, 278.

are the most common issues in the use of hedonic regression analysis in real estate valuation. These basic, most fundamental tests of model rigor address how well the explanatory variables explain the dependent variable and whether the explanatory variables belong in the model. These concepts are as follows:

Model Design

- Proportion of variability explained (R^2 and adjusted R^2)
- The choice of explanatory variables and statistical significance

Model Robustness

- Multicollinearity
- Heteroskedasticity
- Omitted variables
- Sensitivity to outliers

Case Study Example: Using Structural Regression Modeling to Estimate the Impact of a Remediation Program on Property Values in Port Hope, Ontario

In May 2012, a research study was initiated on residential sale prices in the municipality of Port Hope, Ontario, with respect to potential impacts resulting from activities related to the remediation of historic low-level radioactive waste (LLRW) in the community. Under a legal agreement signed in 2001, Canada agreed to undertake a project to remediate the waste and named the project the Port Hope Area Initiative (PHAI).

The purpose of the study was to determine if PHAI activities, such as the signing of the legal agreement, the announcement of remedial plans, or the commencement of remediation activities, had an impact on residential sale prices in the community of Port Hope. These remediation plans included construction of a long-term waste management facility (LTWMF), excavation and dredging activity along the waterfront to remove LLRW, and soil removal and replacement at large-scale sites and individual properties throughout the community. Although remediation could be expected to benefit nearby properties in the long run, the goal of the study was to identify and quantify any potential adverse impacts associated with interference from the remediation activities (e.g., increased truck traffic

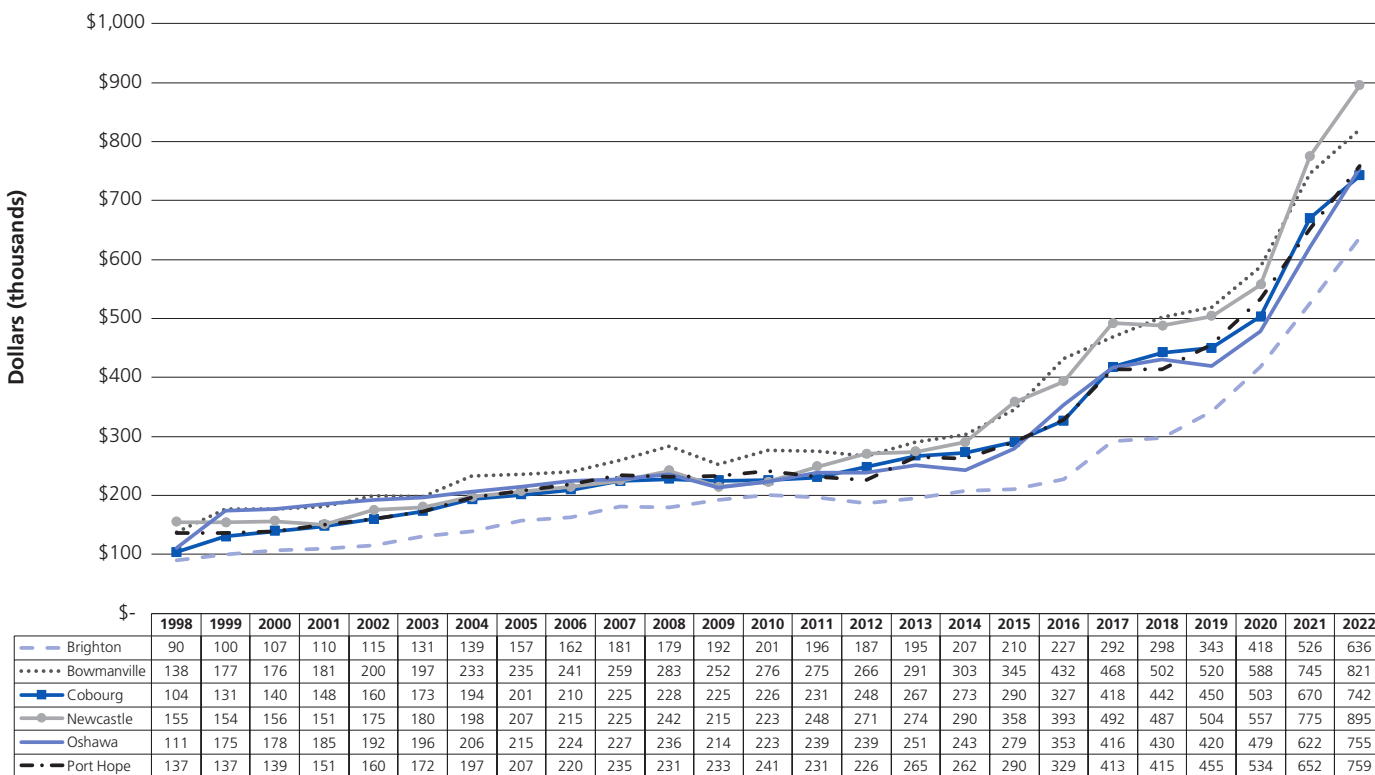
along transportation routes, noise, and the nuisance of soil removal and replacement associated with cleanup activities).

To test price impacts due to location in the area subject to remediation of the historic LLRW, the study used structural regression modeling. Regression analysis provides an efficient and objective analysis of residential sale patterns across the subject area and over time. The study, updated annually or biannually since 2011, analyzed 9,952 sales of single-family residential homes from 1998 through 2022 in Port Hope neighborhoods (subject) and in neighborhoods within five comparable communities unaffected by the remediation of historic LLRW (control). These control neighborhoods are located within the communities of Brighton, Bowmanville, Cobourg, Newcastle, and Oshawa. The control neighborhoods were chosen based upon their comparability to the Port Hope subject neighborhoods and properties. The analysis includes both community-wide and neighborhood-specific models.

The modeling results for the community-wide analysis are visually summarized in Exhibit 1. For the community-wide analysis, sale price is modeled as a function of above-grade floor area, finished basement area, total number of bathrooms, effective age, the presence of multiple stories, total number of fireplaces, total porch points, the presence of central air-conditioning, lot size, the use of nontypical financing (such as seller financing or subsidized loans), year of sale, and municipality in which the property was located. The adjusted R^2 for the community-wide model was 86.4%, indicating that the model explains approximately 86.4% of the variation in sale prices.

Port Hope, along with the control communities, has seen an upward trend in residential sale prices since 2014. As seen in Exhibit 1, sale prices in Port Hope have remained bracketed by sale prices in the control communities over time. There is no pattern of decline or negative price divergence in Port Hope relative to the control communities. All communities, Port Hope included, reached all-time price highs in 2022. The results of the neighborhood-specific analyses were consistent with these findings. The model design considerations and tests for model robustness discussed throughout the following sections were implemented in the development and application of these models.

Exhibit 1 Mean Adjusted Sale Prices by Municipality, 1998 to 2022



Note: The mean adjusted sale prices for each municipality were calculated using only the sales in each delineated control area neighborhood, which were selected based upon their comparability to the Port Hope subject areas. Prices displayed are in thousands of dollars (100 = \$100,000).

Model Design

A model’s design determines its usefulness and applicability to a particular valuation question. When designing a model, an appraiser should consider the nuances of each market under study, including relevant neighborhood boundaries, specific property and sale characteristics, and external influences. Regression models are not “one size fits all,” but instead should be designed with careful consideration of these market and property characteristics. The success of a model developed for one market does not ensure applicability to other market areas.

In the case study example, a series of models were designed to identify adverse pricing divergences between the Port Hope subject neigh-

borhoods and the control area neighborhoods. Consideration was given to neighborhood boundaries and value-driving property and sale characteristics. As a result, the models for each control neighborhood vary slightly to account for the market-specific nuances of each neighborhood.

Proportion of Variability Explained (R^2 or Adjusted R^2)

The R^2 or adjusted R^2 typically represents the most common and primary measure of a model’s explanatory power.¹³ The R^2 measures the percentage of variation in the dependent variable explained by the independent or explanatory variables. For example, if a model of house prices as a function of lot size, living area in square feet,

13. Adjusted R^2 differs from R^2 due to the adjustment for the number of variables used in the model, applying a penalty factor that increases as the number of variables included also increases. *The Appraisal of Real Estate*, 15th ed., explains, “The coefficient of determination, R^2 , can vary from 0 to 1, with 0 indicating no explanatory power whatsoever and 1 indicating perfect explanatory power (i.e., a deterministic model). Adjusted R^2 is useful for comparing multiple competing models with differing sets of independent variables because the measure accounts for the number of explanatory variables in relation to sample size.” (Appendix B: 2)

number of bathrooms, condition, and sale year returns an R^2 of 0.87, the R^2 indicates the explanatory variables explain 87% of the variation in house prices in that market. Adding explanatory variables to the model may artificially increase the R^2 .¹⁴ To account for the artificial increase, the adjusted R^2 adjusts the R^2 value based on the number of explanatory variables included in the model. Therefore, the amount of explanatory power does not increase simply due to the addition of more explanatory variables.

A low R^2 indicates the model does little to explain changes or variation in the variable of interest. Using the previous example, if the sale price model returned an R^2 of 0.34, it indicates the model's explanatory variables explain only 34% of the variation in the dependent variable, *Sale Price*. A model with this low an R^2 indicates the majority of economic drivers in that market are unaccounted for and would typically not be accepted as a defensible representation of house prices or of the contributory value of property and sale characteristics for that local real estate market. On the other hand, a model returning an exceptionally high R^2 (95%–100%) could indicate an overfit model.¹⁵ This type of complication usually arises when the unadjusted R^2 in a regression model contains a significant number of irrelevant explanatory variables. The coefficients associated with those variables likely represent noise, not a real relationship with the dependent variable. As a test, when an overfit model is applied to another dataset or after including additional time periods to the sample dataset, it breaks down and will not adequately represent the new data. Overfit models can be detected using predicted R^2 . “Predicted R^2 is a statistic that measures the ability of a regression model to predict a set of newly presented data....Predicted R^2 reveals if overfitting of the original model exists. The higher the predicted R^2 is, the better the original model is. The lower the value is, the more unsuitable the model is, indicating overfitting.”¹⁶

Generally, analysts seek an R^2 that reaches a

certain threshold. In real estate, acceptable values vary widely across markets and property types. For example, commercial property models typically produce lower R^2 's. What constitutes a minimum R^2 for a reliable model ultimately depends on the data analyzed and purpose of the model—there is no prespecified threshold.¹⁷ As an illustration, the R^2 and adjusted R^2 values reported in recent *Appraisal Journal* articles that involved regression analysis have been tabulated in Exhibit 2. Articles using regression to quantify specific property effects (water views, proximity to a particular amenity or disamenity, and so forth) tended to report adjusted R^2 values largely in the range of 76% to 83%, with the lowest value occurring at 67% and the highest at 83%. On the other hand, studies using regression to quantify market trends (the tendency for properties in an area to appreciate or depreciate across time) typically reported lower R^2 values, in the range of 46% to 61%.

When a regression model is designed to answer a real estate valuation question, the explanatory power (R^2) of a model can often be improved by decreasing the heterogeneity of properties analyzed. When a more homogenous dataset is analyzed, the model can better distinguish the contributory value of the property and sale characteristics (the independent variables). For example, a model analyzing a diverse set of homes covering a large geographic area and with disparate property characteristics will typically have a lower R^2 than a model analyzing more homogenous homes, all else equal. A model analyzing sales of both rural farmland and high-density urban housing will likely be less robust than a model analyzing sales within a planned single-family residential neighborhood. Similarly, models including single-family homes, condominiums, mobile homes, or homes in distinct neighborhoods with disparate value-influencing characteristics will typically have lower explanatory power due to the unique nature of each of the differing property types or markets within a single model.

14. According to *The Appraisal of Real Estate*, 15th ed., “Note that inclusion of any variable, relevant or not, will result in an increase in the coefficient of determination, R^2 . Adjusted R^2 provides a test of whether inclusion of an additional variable adds sufficient explanatory power. When adjusted R^2 does not increase with the addition of another variable, the additional variable is most likely irrelevant.” (Appendix B: 17)

15. See G. Grekousis, *Spatial Analysis Theory and Practice: Describe - Explore - Explain Through GIS* (Cambridge University Press, 2020).

16. Grekousis, *Spatial Analysis Theory and Practice*.

17. *Reference Manual on Scientific Evidence*, 3rd ed., 345.

Exhibit 2 R^2 Values Reported in *Appraisal Journal* Articles Involving Regression Analysis

Appraisal Journal Article Titles	Adjusted R^2	
	Low	High
Topic: Regression models used to quantify specific property effects (structural regressions)		
K. F. Man, and P. P. Mok, "An Empirical Study of the Impacts of an Express Rail Link on Property Prices—Hong Kong Evidence," <i>The Appraisal Journal</i> (Summer 2016): 259–268.	72.9%	73.8%
C. Mothorpe and D. Wyman, "Appraisal of Residential Water View Properties," <i>The Appraisal Journal</i> (Spring 2017): 130–141.	67.0%	72.0%
J. E. Larsen and J. W. Coleman, "Cemetery Proximity and Single-Family House Price," <i>The Appraisal Journal</i> (Winter 2010): 33–49.	71.0%	93.0%
T. O. Jackson and C. Yost-Bremm, "Environmental Risk Premiums and Price Effects in Commercial Real Estate Transactions," <i>The Appraisal Journal</i> (Winter 2018): 48–67.	75.1%	92.6%
S. P. Fraser and M. T. Allen, "Golf Course Design and Real Estate Values: The Impact of Cart Paths on Condominium Prices," <i>The Appraisal Journal</i> (Spring 2017): 96–121.	77.5%	78.4%
O. C. Anderson, C. Yost-Bremm, S. G. Valdez, J. Borrás, and T. Harder, "PFAS Contamination and Residential Property Values: A Study of Five US Sites within the Assessment Stage of the Remediation Lifecycle," <i>The Appraisal Journal</i> (Winter 2022): 26–50.	68.4%	78.7%
J. Laurice and R. Bhattacharya, "Prediction Performance of a Hedonic Pricing Model for Housing," <i>The Appraisal Journal</i> (Spring 2005): 198–209.	79.3%	86.5%
T. Tatos, M. Glic, and T. Luntk, "Property Value Impacts from Transmission Lines, Subtransmission Lines, and Substations," <i>The Appraisal Journal</i> (Summer 2016): 205–230.	89.8%	92.0%
S. C. Bottemiller and M. L. Wolverton, "The Price Effects of HVTLS on Abutting Homes," <i>The Appraisal Journal</i> (Winter 2013): 45–62.	92.9%	93.5%
A. K. Reichert, and H. Y. Liang, "An Economic Analysis of Real Estate Conservation Subdivision Developments," <i>The Appraisal Journal</i> (Summer 2007): 236–245.	68.6%	68.6%
<i>Minimum R^2: 67.0% Maximum R^2: 93.5% Mean Minimum: 76.3% Mean Maximum: 82.8%</i>		
Topic: Regression for measurement of market time trends (trendline regressions)	Low	High
S. Frayn, "Forecasting Commercial Real Estate Appreciation with Commercial Land Sales," <i>The Appraisal Journal</i> (Spring 2014): 133–137.	51.0%	61.0%
R. J. Roddewig, C. T. Brigden, and A. S. Baxendale, "A Pipeline Spill Revisited: How Long Do Impacts on Home Prices Last?," <i>The Appraisal Journal</i> (Winter 2018): 23–47.	46.0%	60.0%
<i>Minimum R^2: 46.0% Maximum R^2: 61.0% Mean Minimum: 48.5% Mean Maximum: 60.5%</i>		

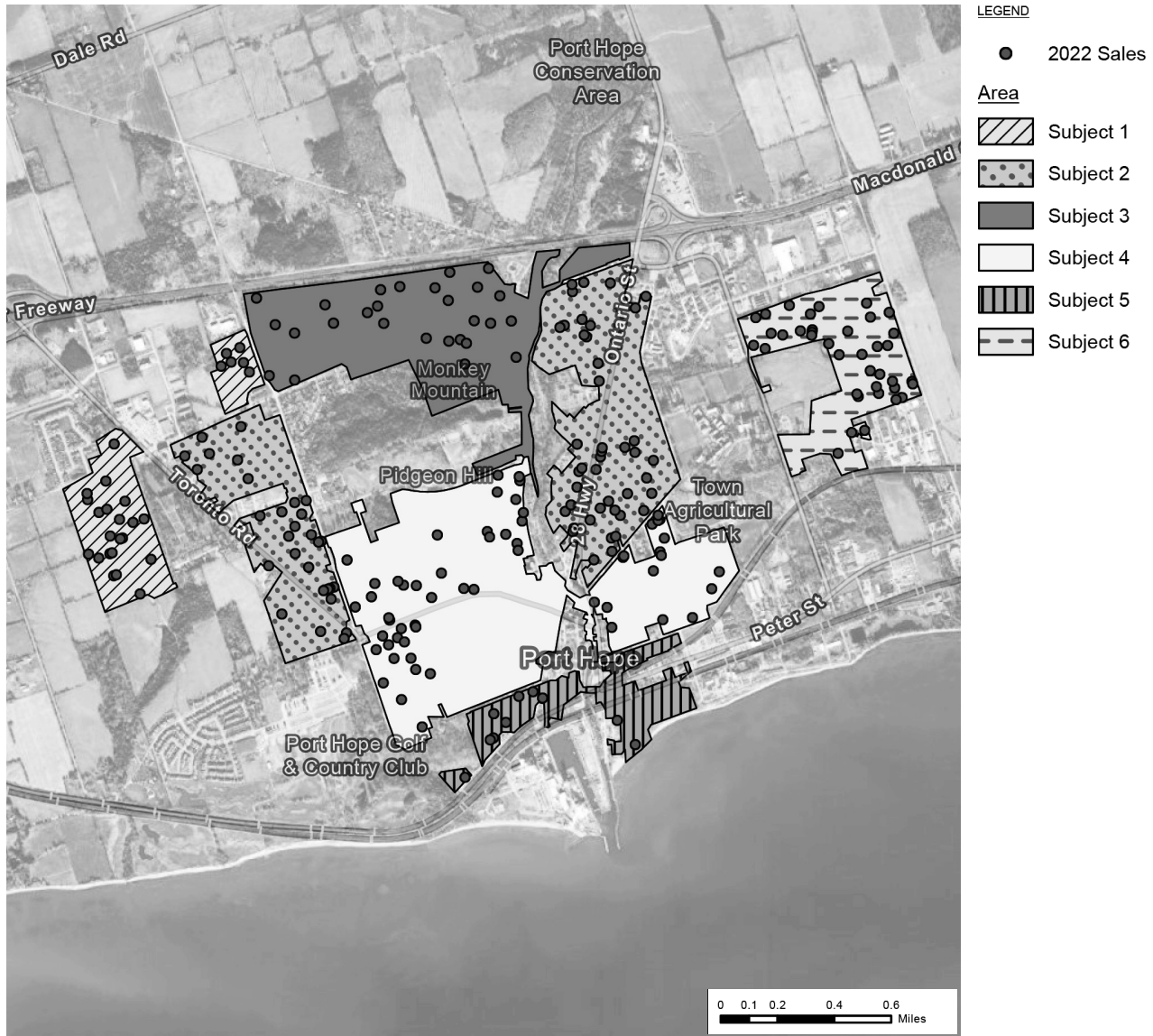
In the case study, subject and control neighborhoods were delineated based on similarity of market and property characteristics. Analyzing these smaller, more homogenous groups of properties individually, in addition to the community-wide comparison, allowed for a more granular and robust modeling of certain areas of Port Hope proximate to certain remediation activities over time. A map of the Port Hope subject neighborhoods is presented in Exhibit 3. The R^2 for both the community-wide and neighborhood models

ranged from 80.6% to 97.5%, meaning the models explained between 80.6% and 97.5% of the variation in house prices.

The Choice of Explanatory Variables and Statistical Significance

Economic theory and data consideration determine the choice of the explanatory variables. Appropriate explanatory variable selection tends to be the most important and tedious task in developing a regression model. Wooldridge states,

Exhibit 3 Map of Port Hope Subject Neighborhoods



Sources: Esri, TomTom, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community, Maxar

“An empirical analysis, by definition, requires data. After data on the relevant variables have been collected, econometric methods estimate the parameters in the econometric model to formally test hypotheses of interest.”¹⁸ Therefore, in many instances, the model and its reliability depend on data quality and availability. If data is unavailable or only poor-quality data is available,

a regression model will not be an appropriate appraisal option.

After the identification of the variables and ensuring the availability of data, the variables must be appropriately specified. “Specification” refers to the form the data takes when entered into the model. Variables can be specified with no change to the data, or they can be transformed

18. Wooldridge, 5.

(e.g., logarithmically, exponentially, lagged, included as a binary or continuous variable).¹⁹ Improper specification affects the model's robustness, potentially yielding biased or inefficient coefficients.

Specifications involving transformations most often include log transformation, either of the dependent variable (e.g., the log of sale price) or the independent variables.²⁰ Transformations like these may both aid interpretation and reduce the impact of outliers or non-normal data.

The appropriateness of an explanatory variable—both the choice to include it and its specification—can be measured by the statistical significance of the coefficient associated with the variable. Statistically significant coefficients indicate a relationship unlikely due to chance (i.e., that the economic effect of the coefficient is real and not a mirage of the data). *P*-values, a measure of the probability of obtaining results as extreme as the observed results of a statistical hypothesis test, measure statistical significance.²¹

In addition to statistical significance, appraisers should also consider the magnitude and sign of the coefficient. A statistically significant coefficient with an extremely small magnitude indicates the variable has minimal effect on the dependent variable. Such a situation indicates the variable is not economically significant. For example, if a model returns a statistically significant coefficient for lot size with a magnitude of 0.0032, it may be concluded that, although statistically significant, the results indicate a lack of economic significance. Therefore, the variable would not likely be included in the final model specification.

Another indication of misspecification, or that a variable may not belong in a model, is when the model returns a coefficient with a sign that does not make economic sense. For example, if a model of property value returns a negative coefficient for lot size (which indicates that a larger lot is worth less than a smaller lot), the appraiser can infer that practically, this does not make sense and, therefore, the model may be suffering from a misspecification issue.²²

Lastly, coefficients should be stable. If small changes (e.g., the addition of sales data, adding or removing an independent variable) to the model cause the coefficients to substantially change their magnitude, sign, or significance, the robustness and quality of the output comes into question. A strong model should return similar results regardless of minor changes imposed on it.

In the case study example, data on sales occurring in the subject neighborhoods in Port Hope and in the comparable control neighborhoods was obtained from 1998 through 2022 from the Municipal Property Assessment Corporation (MPAC). Data included information on potential value-influencing property and sale characteristics that were then tested for statistical significance. If the coefficient associated with the explanatory variable was statistically significant (with a *p*-value of 0.05 or less), it was included in the model. The variables included in each set of neighborhood models differ slightly. The variables were selected based upon the unique property characteristics specific to each subject and control neighborhood being analyzed and the statistical significance of each variable for those specific areas.

19. Logarithmic specifications of independent variables involve taking the natural log of the variables before estimating the regression. This can reduce the impact of extreme values of the variable under study and capture the economic principle of decreasing marginal utility (e.g., the dollar benefit of an additional square foot of space decreases, as a house becomes larger). Lagged variables mean the inclusion of observational data that is not contemporaneous with the value of the independent variable data and is less commonly used in hedonic real estate regressions. Binary (or more generally, categorical) variables capture category-like effects, such as the year of sale, or the external influence of being located in a particular neighborhood or school district.

20. When variables included in the model contain negative values, a box-cox transformation may be used instead of a log transformation. See, for example, R. Davidson and J. Mackinnon, "Testing Linear and Loglinear Regressions against Box-Cox Alternatives," *The Canadian Journal of Economics* (August 1985): 499–517.

21. See, for example, the discussion of *p*-values from *Reference Manual on Scientific Evidence*, 3rd ed., 250: "Large *p*-values indicate that a disparity can easily be explained by the play of chance: The data fall within the range likely to be produced by chance variation. On the other hand, if *p* is very small, something other than chance must be involved." What constitutes a "small" *p*-value for the purpose of determining statistical significance often depends on the data and models used, convention within the discipline, and a careful assessment of the range of possible effects. For greater detail, see Greenland et al., "Statistical Tests, *P*-values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology* 31, no. 4 (2016): 337–350, and *The Appraisal of Real Estate*, 15th ed., Appendix B.

22. Counterintuitively signed coefficients can also be an indication of multicollinearity. See the discussion of model robustness below.

Both sale price and the natural log of sale price were modeled, with the latter included to minimize the impact of potential outlier (i.e., highly unusual) sales. The use of the logarithmic models also provides an additional beneficial interpretation for the effects measured. The logarithmic transformation of the property sale price can be interpreted as the percentage change in the price of the home, whereas linear models estimate the simple dollar difference.

Model Robustness

In addition to the basic tests of model fit, several robustness tests should be run to ensure a model's reliability.²³ Multicollinearity, heteroskedasticity, omitted variable bias, and sensitivity to outliers tend to be the issues that plague real estate modeling most often. Appraisers must be aware of these issues and test for them when using regression analysis and interpreting its output.

Multicollinearity

A high degree of correlation between two or more explanatory variables (i.e., the explanatory variables are strongly related) indicates that multicollinearity exists. The explanatory variables included in a model should be independent. If they are not, "multicollinearity does seriously affect structural interpretation of a model's coefficients."²⁴ A strong relationship between two or more of the variables make it difficult, if not impossible, for the model to isolate each variable's individual effect. Multicollinearity causes biased coefficients of the correlated variables, making them unstable and difficult to interpret.²⁵ For example, if the existence and percentage of property value diminution from a disamenity is being estimated, the coefficient explaining the effect will be inaccurate if multicollinearity is present and acting upon this variable. Because hedonic regression ultimately relies on correct interpretation of the model's coefficients, the higher the degree of multicollinearity, the more statistically unreliable any output from that analysis will be.²⁶

The effect of multicollinearity on coefficients is difficult to disentangle. The presence or degree of multicollinearity cannot be identified prior to model estimation nor is the direction of bias on the coefficient consistent. Therefore, the presence of multicollinearity cannot be ruled out until a model is specified and multicollinearity is tested for. The nature of multicollinearity also does not allow for interpreting the output as a "best case" or "worse case" scenario for a particular measured effect. Another malignant side effect of this variable entanglement is that the coefficients become sensitive to small changes in the model specification. That is, the magnitude and signs of the coefficients change easily and often disproportionately when even small adjustments are made to the model.

Consider a concrete example of multicollinearity in which a single market area contains multiple similar contamination sources. Such a market area is not uncommon in industrial areas, including areas that have been converted to residential use. If the model includes multiple contamination variables representing different potential contamination sites, the model may have difficulty differentiating the effect on property values from each individual site. This is a result of the contamination variables likely having a strong correlation. As a result, the model will produce and assign erroneous estimates of the effects from any one of the individual sites. If this issue is not addressed, the modeler can draw erroneous conclusions about which environmental condition is producing the measured effect and by how much.

Exhibit 4 shows a hypothetical example illustrating the effects of multicollinearity on the reliability of statistical output. The table presents output from two models that estimate an adjustment factor for a two-story home versus a single-story home. Model A includes variables correlated to the effect of a second floor. (For simplicity, their effects are not shown here.) Model A estimates that market participants discount a multistory configuration by \$12,000 when compared to an otherwise similar single-story home. Moreover,

23. From *Reference Manual on Scientific Evidence*, 3rd ed., 311: "[robustness]: Seeing the extent to which the results are sensitive to changes in the underlying assumptions of the regression model."

24. *The Appraisal of Real Estate*, 15th ed., Appendix B: 15.

25. *The Appraisal of Real Estate*, 15th ed., Appendix B: 15, "[T]he independent variables share explanatory power and consequently the coefficients on the correlated independent variables are biased."

26. *Reference Manual on Scientific Evidence*, 3rd ed., 324.

Exhibit 4 A Hypothetical Example of Multicollinearity Effects on Coefficient Estimates

	Model A. Effect of a Second Floor (Multicollinearity Present)	Model B. Effect of Second Floor (Multicollinearity not Present)
Model Measured Effect	−\$12,000	−\$2,000
P-value (Statistical Significance)	<0.01	0.42

the effect is significant, with a two-sided p -value of <0.01. (A p -value below the conventional threshold of 0.05 is typically considered to be statistically significant.) An appraiser who failed to recognize and correct for the multicollinearity in this model might erroneously conclude that multi-story homes in this market sell for less than single-story homes with otherwise similar features.

In Exhibit 4, the model presented in the right column, Model B, removes the collinear variables. The output from Model B indicates no statistically significant negative effect from the presence of a second story, when compared to an otherwise similar single-story home. While the direction of the effect is still negative (a \$2,000 discount), it is no longer statistically significant, with a p -value of 0.42. The high p -value indicates any multistory effect is likely due to chance. Model B corrects for multicollinearity and indicates no detrimental effect attributable to a multi-level configuration.

Correlation matrices and variance inflation factors (VIFs) test for multicollinearity.²⁷ VIFs are most commonly used. They measure the amount each coefficient contributes to the possible total amount of multicollinearity. Scores for VIFs range in value, but values below 10 generally indicate no cause for concern for multicollinearity.²⁸ According to the textbook *Applied Linear Statistical Models*, “A maximum VIF value in excess of 10 is frequently taken as an indication that multicol-

linearity may be unduly influencing the least squares estimates.”²⁹ In addition, Wooldridge writes, “If VIF is above 10 then we conclude that multicollinearity is a ‘problem’ for estimating,” although he goes on to say that “a VIF above 10 does not mean that the [uncertainty of estimates] are too large to be useful.”³⁰

Moderate levels of multicollinearity may not be as problematic. However, the extent to which its presence may impact the results depends on whether it affects the main structural variables of interest. Analysis of pairwise correlation can help appraisers infer which variables are most likely to be collinear, but this is not always a fail-safe approach because linear dependence between variables can exist among more than two variables at one time.

It is imperative to test for and attempt to reduce the impact of multicollinearity. It can affect the values and signs of the explanatory variables, as well as the measure of statistical significance—whether the effect exists or is due to random chance. Multicollinearity can only be identified after data is collected, the model is specified, and regression analysis is conducted. Once multicollinearity is identified, its effect is inconsistent and proves difficult to correct for. Simply amassing a larger dataset, for example, will not solve the problem.³¹ If multicollinearity is identified, an appraiser may need to revisit the selection of explanatory variables and remove

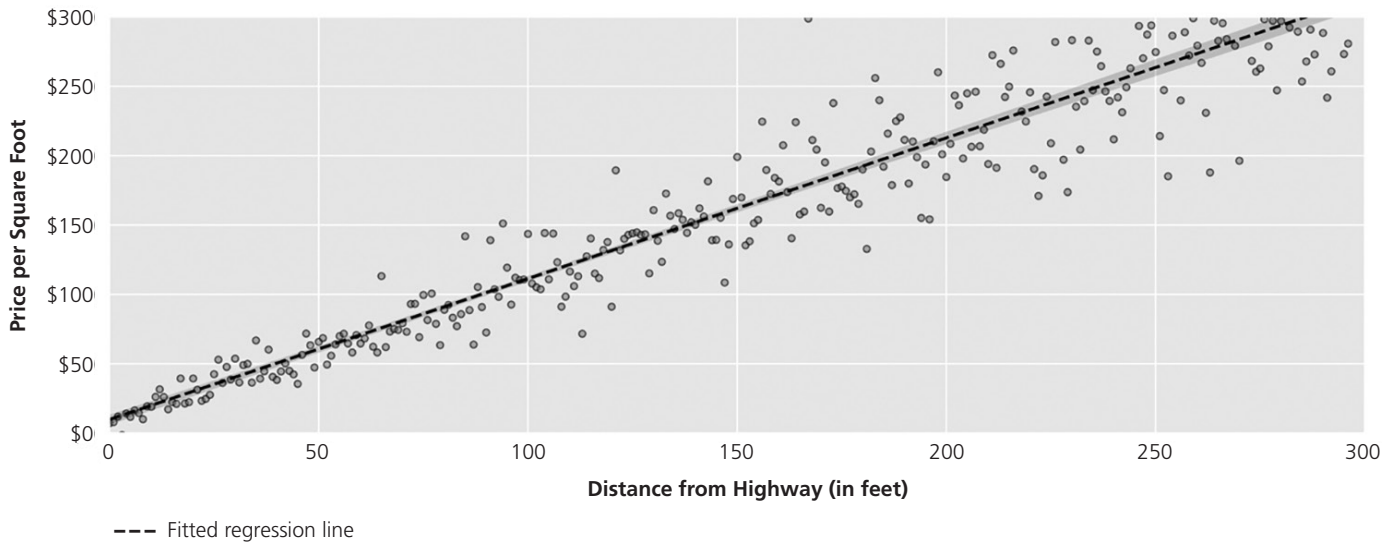
27. See, for example, *The Appraisal of Real Estate*, 15th ed., Appendix B: 15: “Investigation of the existence of multicollinearity includes analysis of a matrix of independent variable correlation and an examination of regression model multicollinearity diagnostics including variance inflation factors (VIFs).”

28. *The Appraisal of Real Estate*, 15th ed., Appendix B: 15.

29. Kutner, Nachtsheim, Neter, and Li, *Applied Linear Statistical Models*, 5th ed., Chapter 10: Building the Regression Model II: Diagnostics (2005), 409.

30. Wooldridge, 86. The paraphrasing is of mathematical formulae.

31. “A relatively small sample, or even a large sample with substantial multicollinearity, may not provide sufficient information for the expert to determine whether there is a relationship,” according to *Reference Manual on Scientific Evidence*, 3rd ed., 325.

Exhibit 5 Illustration of Potential Heteroskedasticity

collinear variables that are affecting the structural variables of interest.

In the case study example, multicollinearity was tested for using VIFs. Multicollinearity was not an issue in the Port Hope models.

Heteroskedasticity

Heteroskedasticity is a different, yet related, issue to multicollinearity. While multicollinearity creates biased estimated coefficients, heteroskedasticity affects the statistical reliability of estimates. Heteroskedasticity masks the significance of otherwise significant explanatory variables. All estimates from a regression model are measured with some margin of error, by which the uncertainty about the precise value is quantified. An “estimate,” in other words, is just that: a deduction about a particular effect, accurate only within a certain range. If heteroskedasticity is present and unaccounted for, the true margin of error—that range—will be different than what is measured and shown in the statistical output. Heteroskedasticity will not alter the value of the coefficient but alters the reliability and may result in a coefficient being interpreted as significant when in fact it is not or vice versa.

Like multicollinearity, heteroskedasticity cannot be assumed away or identified prior to estima-

tion. In fact, a model used to describe the variation in sale prices in one market area may exhibit no heteroskedasticity but, when applied to a different real estate market, may exhibit high levels of heteroskedasticity, thus rendering the model’s findings unreliable.

Heteroskedasticity occurs when the mean and variance of the model’s error term is not steady across the range of one or more explanatory variables, graphically demonstrated in Exhibit 5. The illustration displays a fitted regression line (i.e., the values of each home that are predicted by the model), which corresponds to a regression model’s estimate of the relationship between a single explanatory variable (a property’s distance from a major highway) and the dependent variable (a property’s sale price per square foot).³²

As can be seen in Exhibit 5, sale price variability increases as distance from a highway increases. The evenly balanced variability above and below the regression line illustrates how heteroskedasticity does not affect the model’s estimate of the explanatory variable on sale price (i.e., the coefficient). Instead, heteroskedasticity only affects the reliability and its generalizability (its statistical significance, or probability value). Heteroskedasticity often under- or overstates the statistical significance of variables, making

32. This is a simplified example, for illustration only. An alternative (but less intuitive) approach is to plot model errors against the fitted regression line produced by the model. See *The Appraisal of Real Estate*, 15th ed., Appendix B: 14.

the model more likely to identify a statistical effect when one does not exist, or vice versa.

A relatively simple solution to address heteroskedasticity is to reestimate the model using what is known as “robust” standard errors. A robust standard error incorporates non-constant variance into the formula of the standard error and therefore accounts for heteroskedasticity in the model.

In the case study example, heteroskedasticity was identified in several iterations of the models. To address this, all models were reestimated using robust standard errors.

Omitted Variables

When economically important variables are omitted from a model, a portion of the economic effects from the missing variables get falsely attributed to the economic effects of the included variables.³³ This has clear implications for inferring cause-and-effect relationships using regression. If an important variable is omitted from the regression model, the economic effects of the remaining variables may be under- or over-stated. The *Reference Manual on Scientific Evidence* states, “Failure to include a major explanatory variable that is correlated with the variable of interest in a regression model may cause an included variable to be credited with an effect that is actually caused by the excluded variable. In general, omitted variables that are correlated with the dependent variable reduce the probative value of the regression analysis.”³⁴ Failure to include the right variables may result in a model that falsely identifies a contributory effect for an included variable when there is not one, or vice versa.

Double-counting in adjustments is an issue appraisers are well aware of. While regression models can help isolate and quantify various contributory and detrimental effects on value—from bedrooms, to size, to many more potential influences—it is not a panacea that is immune from traditional appraisal considerations. Any potential for omitted variables needs to be thoroughly explored in a model before the model’s coefficients can be reliably interpreted.

Uncovering precisely how the presence of omitted variables may impact the reliability of a model is a complex process. In general, the higher the degree of correlation between the omitted variables and the included variables, the more bias present in the effects measured by the model. The *Reference Manual on Scientific Evidence* states, “The importance of omitting a relevant variable depends on the strength of the relationship between the omitted variable and the dependent variable and the strength of the correlation between the omitted variable and the explanatory variables of interest. Other things being equal, the greater the correlation between the omitted variable and the variable of interest, the greater the bias caused by the omission. As a result, the omission of an important variable may lead to inferences made from regression analyses that do not assist the trier of fact.”³⁵

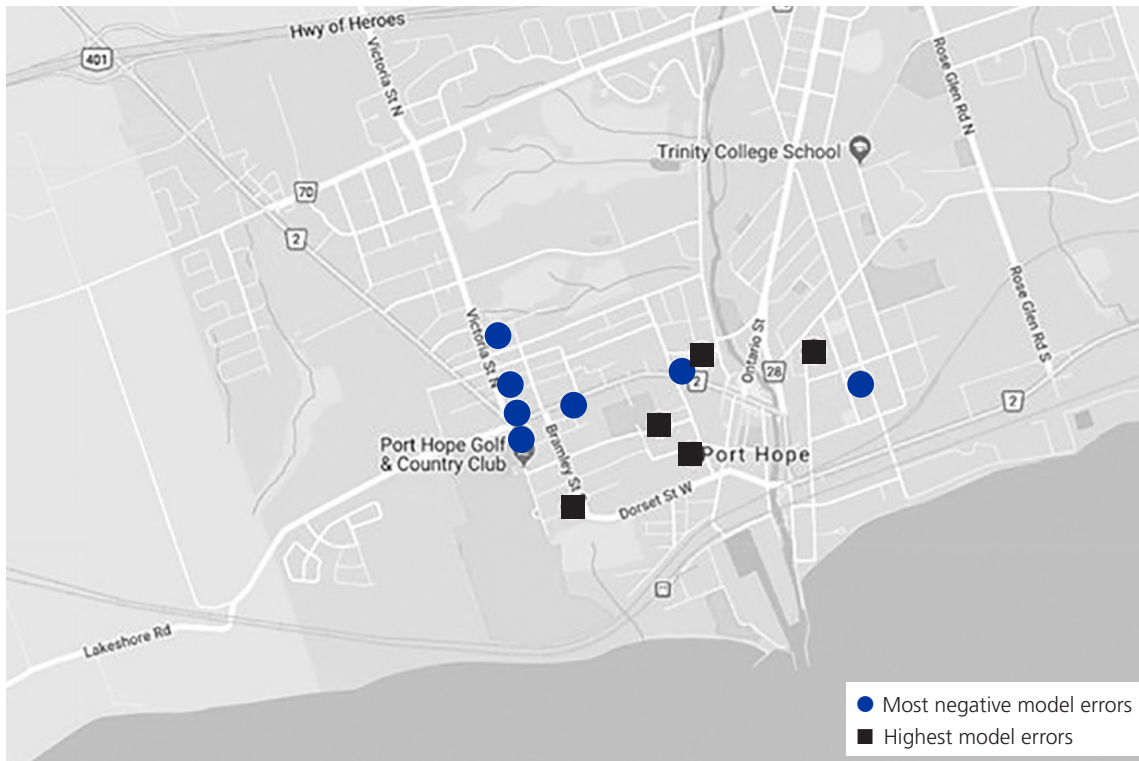
Without including an omitted variable (or variables), it is impossible to know by how much or in what direction the misstatement has occurred. Appraisers must acknowledge this potential and present evidence showing they have considered whether variables included in their theoretical formula sufficiently avoid omitted variable bias.

What constitutes an omitted variable problem depends on the specifics of the relationship being modeled, in this case, the local real estate market. There is no one formula (or econometric specification) that can be used as a one-size-fits-all approach. Each market has unique considerations impacting its real estate values. For example, the presence of a pool likely contributes little to no value in some markets but could be a significant amenity in others. Therefore, omitting data on the presence of a pool may present little issue in one market but produce substantially inaccurate estimates of property value effects in other areas. Because of the unique nature of real estate, no single model is generally applicable to all real estate markets. As with other valuation tools, in-depth knowledge and research of the market and properties under study is crucial for appropriate variable selection and reliable output from a regression model.

33. *The Appraisal of Real Estate*, 15th ed., Appendix B: 11.

34. *Reference Manual on Scientific Evidence*, 3rd ed., 314.

35. *Reference Manual on Scientific Evidence*, 3rd ed., 314.

Exhibit 6 Geospatial Mapping of Residuals in Subject Area 3 in 2018

In the case study example, a statistically significant negative price divergence was preliminarily identified in two of the subject area neighborhoods from 2017 to 2018. The mean adjusted sale price in Subject Areas 3 and 4 experienced a decrease relative to the control areas during this time frame. This time frame was not concurrent with any remedial activities within or proximate to those specific neighborhoods. To determine the source of the negative price divergence, further investigations into the sales in Subject Areas 3 and 4 in Port Hope were conducted.

As part of this investigation, the model residuals (i.e., errors) were geospatially mapped to visually identify any potential geographic patterns. The geospatial mapping could indicate whether the existing model was systematically under- or over-predicting property sale prices in a concentrated geographic area. Exhibit 6 displays the mapped model residuals for property sales in Subject Area 3 in 2018. The blue-circle values represent the approximately 16% of the sample with the most negative model errors (and thus had lower prices than what the model would otherwise predict), while the black-square values

represent approximately 16% of the highest model errors (and thus have higher actual sale prices than what the model would predict).

The geospatial mapping indicated that some of the 2018 property sales with unusually negative residuals appeared to cluster next to high-traffic thoroughfares. These negative residuals indicated that sale prices for these particular properties were lower than what the statistical model would have predicted otherwise. Therefore, since a variable for proximity to high-traffic thoroughfares was omitted from the original model, the model erroneously attributed the negative change in sale price to location in the subject area when compared to the corresponding control areas, when in fact it was driven by the greater number of 2018 subject sales located along a high-traffic thoroughfare.

Information was used in the dataset received from MPAC to identify properties in close proximity to light-, heavy-, or medium-traffic thoroughfares. Two binary variables were constructed and added to the models, one indicating proximity to “light or medium” traffic, and the second indicating proximity to “heavy” traffic. Results

indicated that exposure to “light or medium” traffic had no statistical effect on pricing for Area 3. On the other hand, exposure to “heavy” traffic was associated with a statistically significant reduction in property value. For example, the linear model for Subject Area 3 indicated a price reduction of $-\$15,547$ for proximate properties. After controlling for heavy traffic, three of the four model specifications for Area 3 (linear and logarithmic, both with and without outliers) no longer showed negative sale price effects for the subject areas. Area 4 models showed similar results, also indicating a detrimental effect from heavy-traffic proximity once that variable was added to the model.

Sensitivity to Outliers

The presence of outliers in the data affects the reliability of the results from a regression model. In practice, the presence or extent of an outlier problem cannot be ascertained by merely looking at a regression model formula. Actual data needs to be analyzed and rigorously tested in order to determine whether a particular regression model can be used to reliably estimate property value effects in a specific market. At best, a model sensitive to outliers will return clearly erroneous results. At worst, it produces results that appear to provide usable insights but are in fact driven by only a few unusual data points.

Exhibit 7 contains two graphs depicting the square footage and sale price of hypothetical homes. Exhibit 7-1 includes just three additional houses selling for unusually low prices. These are examples of outliers. The illustration shows how just a few outlier sales can produce erroneous estimates in a regression analysis.³⁶ The graph in 7-1 (depicting the model with the outliers included in the dataset) calculates an increase in value of \$42 for each additional square foot of living area. The graph in 7-2 (depicting the model correcting for outliers) calculates an increase in value of \$63.20 for each additional square foot of living area. In other words, removing these three outliers results

in an increase of the estimated contributory value of an additional square foot by over 50%. ($63.20/42 = 0.5047$).

In the case study example, models were run both with and without outliers removed to ensure that sales with abnormally high or low prices were not biasing or influencing the results of the analysis. Outliers or sales with a price two standard deviations above or below the mean were removed. Results were generally consistent for the models with and without outliers removed, in both the linear and logarithmic specifications.

Conclusion

When employing regression analysis to understand property value, appraisers must be prepared to thoroughly test the validity and robustness of a regression model when applied to a specific valuation problem in a specific market area. A regression model is not a black box capable of answering any question or solving any valuation problem, and appraisers looking to use the tool should engage expert advice or collaboration if they are unfamiliar with the method.

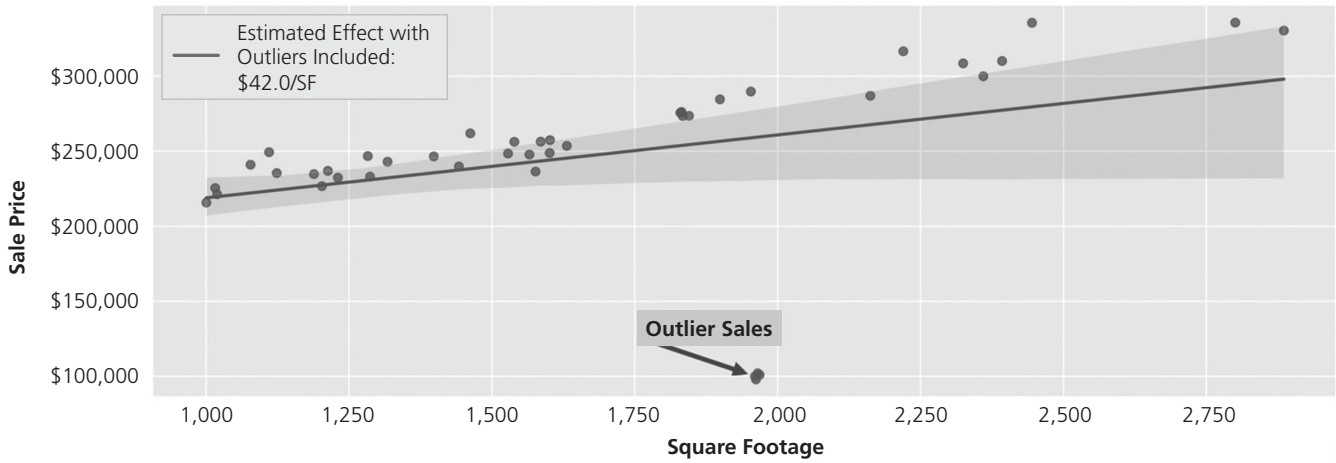
As discussed, the success of a regression model depends on data availability, data quality, proper variable and model specification, and a number of other robustness factors. Many issues affecting real estate models cannot be identified until after a model is run and results are generated for interpretation and scrutiny. As a result, it should never be assumed that a robust model can be developed for any real estate market without going through rigorous development and testing discussed herein. Regression models, when appropriately designed and implemented, can be a powerful tool for identifying various real estate effects. Appraisers should strive to ensure both the quality of the results produced by this tool and their appropriate interpretation and application.

CONTINUED >

36. The smaller the sample size, the greater the influence even just one outlier can have on the overall results.

Exhibit 7 Illustration of the Effects of Outliers

7-1. With a Few Outliers (100 Sales): \$42 per Square Foot



7-2. The Same Data with Three Outliers Removed (97 Sales): \$63.20 per Square Foot (50.4% Increase)



About the Authors

Erin Kiella, PhD, is executive vice president and consultant at Real Property Analytics Inc. Kiella has been with Real Property Analytics since 2015. Her expertise is in complex real estate valuation techniques used to quantify potential property value diminution from detrimental conditions, including environmental contamination or alleged contamination from on-site and off-site sources. She has expertise in statistical modeling and econometrics. Kiella has provided litigation support involving the development of damage and rebuttal opinions in class action and mass tort litigation cases throughout the United States, both at the certification and merits stages. Kiella was formerly an assistant research economist with the Real Estate Center at Texas A&M University, where her research focused on rural land market trends, agricultural lending, and estimating econometric models forecasting rural land prices in Texas, Alabama, Mississippi, and Louisiana. Before joining the Real Estate Center in January 2018, Kiella was a strategy consultant with California-based The Wonderful Company, research assistant with the Agricultural and Food Policy Center at Texas A&M University, and consultant with the Federal Reserve Bank of Chicago. She has lectured several courses at Texas A&M University. Kiella has a PhD in agricultural economics from Texas A&M University and a BBA in finance and economics from Loyola University in Chicago, with honors. She is also a member of the American Society of Farm Managers and Rural Appraisers. **Contact: erin@rpa-inc.com**

Jennifer Pitts, MAI, CRE, is the president of Real Property Analytics Inc. and has over fifteen years of experience in real estate consulting and appraisal throughout the United States and Canada. She specializes in analyzing complex valuation issues, including the valuation of properties impacted by environmental contamination or other disamenities, and has provided expert testimony on litigation matters before federal and state courts. These matters involved the impacts of soil, groundwater, airborne and surface water contamination, and alleged contamination on property values; real estate issues related to proposed environmental class actions; the impacts of high-voltage electric transmission lines on property values; the valuation and highest and best use of properties subject to eminent domain; and real estate development feasibility. She has familiarity and experience with specialized valuation methods used in these types of assignments, including econometric and statistical modeling, paired sales analysis, and case study research. Pitts is a graduate of Texas A&M University with a master's degree in land economics and real estate and a bachelor's degree (summa cum laude) in finance. She is a designated member of the Appraisal Institute (MAI) and a Counselor of Real Estate (CRE)—a professional designation that is awarded by invitation only to a select number of professionals recognized for their expertise, experience, and ethics in providing real estate counseling and advisory services. She is a state-certified general real estate appraiser in Texas and other states. She has coauthored several articles for *The Appraisal Journal* and *The Journal of Real Estate Literature*. **Contact: jennifer@rpa-inc.com**

Christopher Yost-Bremm, PhD, is an assistant professor of finance at San Francisco State University. Yost-Bremm has been with Real Property Analytics Inc. since 2014, developing and critiquing statistical methodologies involving real property on behalf of numerous firms and individuals. He has significant experience in analyzing the impacts of environmental contamination, particularly under class action or mass tort claims (at both the certification and merit stages). In addition, Yost-Bremm has provided analytic valuation services for mining and other industrial properties and has analyzed numerous other commercial and residential property types under complex economic situations, such as low-income housing tax credits, property diminution in insurance claims, transferable development rights, and partial takings in eminent domain, among other matters. In addition, through the use of complex econometric models, he has studied the impacts of environmental contamination on income capitalization rate risk premiums and sale prices for commercial properties in Southern California and coauthored an article on this topic. He has published real estate and financial valuation work in numerous academic journals, including *The Journal of Behavioral Finance*, *Cities*, *The North American Journal of Economics and Finance*, *The Review of Behavioral Finance*, and *The Journal of Computer Information Systems*. He received the Richard U. Ratcliff Award from the Appraisal Institute, presented annually for the most outstanding original article by an academic author published in *The Appraisal Journal*, for his regression study on commercial property values amid environmental contamination. Yost-Bremm holds a PhD in finance from Texas A&M University, an MBA from California State University (with distinction), and undergraduate degrees in management and international economics (with honors). He is a state-certified general real estate appraiser in California. **Contact: chris@rpa-inc.com**

Additional Resources

Suggested by the Y.T. and Louise Lee Lum Library

Appraisal Institute

- **Education**
 - *Quantitative Analysis*
 - *Real Estate Finance, Statistics, and Valuation Modeling*
- **Lum Library Knowledge Base information compilation [Login required]**

Appraisal Practice: Data, statistics, and statistical analysis—regression analysis
- **Publications**
 - *The Appraisal of Real Estate*
 - *An Introduction to Statistics for Appraisers*
 - *Practical Applications in Appraisal Valuation Modeling*
 - *Valuation by Comparison*