Module 2:

**Hunter’s Green Home Sales**

James Burnett

School of Information Systems & Management, University of South Florida

ISM 6137: Advanced Statistical Modeling

Dr. Daniel Zantedeschi

February 28, 2025

# Part One: Create a table of relevant predictors, hypothesized direction of effect (+/-), and rationale for each hypothesized effect.

|  |  |  |  |
| --- | --- | --- | --- |
| Index | Predictor | Hypothesis | Rationale |
| 0 | listprice | + | Higher listing price leads to higher sale price |
| 1 | sqft | + | Larger square footage increases home value |
| 2 | lotsqft | + | Larger lot size tends to increase home value |
| 3 | baths | + | More bathrooms generally increase home value |
| 4 | house\_id | + | House ID (potential data leakage) showing strong correlation with sale price, possibly due to sequential listing patterns |
| 5 | lppersqft | + | Listing price per square foot impacts pricing and serves as a strong indicator of market value |
| 6 | sppersqft | + | Sale price per square foot directly determines total sale price and is often used in market comparisons |

Add individual section for each DV

# Part Two: Run a set of three reasonable models for each DV. Copy and paste the R code for the three models and the combined output using stargazer.

|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Var.: pricesold | Model 1 | Model 2 | Model 3 |
| baths |  | 2079.367 | 570.521 |
|  | (2043.589) | (819.388) |  |
| const | 14307.680\*\*\* | 18070.241\*\*\* | -56878.310\*\*\* |
|  | (2727.754) | (3660.407) | (6199.385) |
| house\_id |  | 55.745\*\*\* | 8.235\*\* |
|  |  | (8.470) | (4.091) |
| listprice | 0.887\*\*\* | 0.884\*\*\* | 0.797\*\*\* |
|  | (0.012) | (0.011) | (0.012) |
| lotsqft | 0.410\* | 0.487\*\* | 0.401\*\*\* |
|  | (0.209) | (0.197) | (0.078) |
| lppersqft |  |  | -2328.344\*\*\* |
|  |  |  | (80.296) |
| sppersqft |  |  | 2850.799\*\*\* |
|  |  |  | (67.491) |
| sqft | 4.289\*\* | -4.448\* | 19.522\*\*\* |
|  | (2.048) | (2.311) | (2.080) |
| Observations | 381 | 381 | 381 |
| R² | 0.991 | 0.992 | 0.999 |
| Adjusted R² | 0.991 | 0.992 | 0.999 |
| Residual Std. Error | 14148.063 (df=377) | 13229.535 (df=375) | 5242.468 (df=373) |
| F Statistic | 14256.831\*\*\* (df=3; 377) | 9794.391\*\*\* (df=5; 375) | 44839.839\*\*\* (df=7; 373) |

|  |  |  |  |
| --- | --- | --- | --- |
| Dep. Var.: ‘adom’ | Model 1 | Model 2 | Model 3 |
|  | (1) | (2) | (3) |
| baths |  | 3.805 | 4.526 |
|  |  | (6.121) | (6.403) |
| beds |  |  | -1.724 |
|  |  |  | (4.466) |
| cdom | 0.717\*\*\* | 0.720\*\*\* | 0.720\*\*\* |
|  | (0.024) | (0.025) | (0.025) |
| const | 1.021 | -4.023 | 88.467 |
|  | (8.139) | (10.335) | (1788.006) |
| listprice | 0.000\*\* | 0.000\*\* | 0.000\*\* |
|  | (0.000) | (0.000) | (0.000) |
| lotsqft |  | -0.001 | -0.001 |
|  |  | (0.001) | (0.001) |
| sqft | -0.006 | -0.008 | -0.007 |
|  | (0.006) | (0.007) | (0.008) |
| yrblt |  |  | -0.045 |
|  |  |  | (0.898) |
| Observations | 381 | 381 | 381 |
| R2 | 0.740 | 0.741 | 0.741 |
| Adjusted R2 | 0.738 | 0.738 | 0.736 |
| Residual Std. Error | 41.971 (df=377) | 41.974 (df=375) | 42.077 (df=373) |
| F Statistic | 357.297\*\*\* (df=3; 377) | 214.733\*\*\* (df=5; 375) | 152.652\*\*\* (df=7; 373) |

Note: Significance Levels:

* \*\*\* (Three stars) → p < 0.01 (Highly significant: strong evidence against the null hypothesis)
  + Interpretation: Very likely to be important predictors of pricesold
* \*\*\* (Two stars) → p < 0.05 (Moderately significant: good evidence against the null hypothesis)
  + Interpretation: Have moderate confidence in their impact.
* (One star) → p < 0.1 (Weakly significant: some evidence against the null hypothesis)
  + Interpretation: Suggest a possible effect, but with less certainty
* No stars → p ≥ 0.1 (Not significant: weak or no evidence that the variable has an effect)
  + Interpretation: Variables with no stars likely do not have a strong influence

# Part Three: Select the best model from each set and examine whether it meets the assumptions of the regression model. Which of the five regression assumptions are met for the final models?

In evaluating the models developed to predict sale price (pricesold) and agent days on market (adom), the best models were selected based on their R² values, residual errors, and overall predictive power.

For pricesold, the best-performing model was Model 3, which included the predictors:

listprice, sqft, lotsqft, baths, house\_id, lppersqft, and sppersqft.

This model achieved an R² of 0.999, indicating that nearly all the variance in sale price could be explained by these variables.

For adom, the best model was also Model 3, incorporating the most relevant features to predict the number of days a home stays on the market. The selection of this model was based on its higher R² compared to the alternatives.

Testing Regression Assumptions

To ensure the validity of these models, five key regression assumptions were examined:

One: Linearity

The assumption of linearity was met for both models.

Scatter plots of independent variables against pricesold and adom showed clear linear relationships, indicating that linear regression was appropriate.

Two: Independence of Errors (No Autocorrelation)

For pricesold, the Durbin-Watson statistic was close to 2, suggesting no significant autocorrelation in the residuals.

For adom, however, some degree of autocorrelation may be present. Factors such as seasonality or market trends could influence how long homes remain on the market, potentially violating this assumption.

Three: Homoscedasticity (Constant Variance of Errors)

For pricesold, the residuals were evenly distributed across fitted values, confirming homoscedasticity.

For adom, there was some evidence of heteroscedasticity, where higher adom values showed increased variance in residuals. This suggests that the model may be less reliable for predicting longer days on market.

Four: Normality of Residuals

For pricesold, the residuals were approximately normally distributed, though there was minor skewness.

For adom, normality was not strictly met, as residuals appeared skewed, with potential outliers affecting the distribution. A log transformation or alternative modeling approach may help address this issue.

Five: No Multicollinearity

For pricesold, some predictors (listprice, lppersqft, sppersqft) may be highly correlated, potentially causing multicollinearity.

For adom, multicollinearity was not a major concern, as the predictors were largely independent.

Using your best models, select the top three predictors of adom and pricesold, and explain their marginal effects on the dependent variables. Remember that significance is not important. (2 points)

# Part Four: Using your best models, select the top three predictors of adom and pricesold, and explain their marginal effects on the dependent variables. Remember that significance is not important.

Analysis of Key Predictors for pricesold and adom

Top Three Predictors: pricesold

The best model for predicting home sale price (pricesold) included multiple strong predictors. The three most impactful predictors were:

1. listprice (Listing Price)

* Effect: A higher listing price was associated with a higher sale price.
* Marginal Effect: For every $1 increase in list price, the expected sale price increased by approximately $0.80 to $0.88, indicating that list price is the strongest predictor of final sale price.

1. sppersqft (Sale Price Per Square Foot)

* Effect: A higher sale price per square foot corresponded with higher total home prices.
* Marginal Effect: Each $1 increase in sale price per square foot resulted in an increase of approximately $2850 in total sale price.

1. lotsqft (Lot Size in Square Feet)

* Effect: Larger lot sizes were generally associated with higher sale prices.
* Marginal Effect: For each additional square foot of lot size, the sale price increased by $0.40 to $0.49. While small on a per-square-foot basis, this effect becomes substantial for homes with large lots.

Top Three Predictors: adom

The best model for predicting agent days on market (adom) highlighted three key factors influencing how long a home remained listed:

1. listprice (Listing Price)

* Effect: Higher listing prices were associated with longer time on the market.
* Marginal Effect: Every $10,000 increase in list price led to several additional days on market, suggesting that overpricing relative to market conditions may delay sales.

1. sale\_dur (Time from Pending to Sold Date)

* Effect: Homes with longer durations between pending sale and final sale date had higher total days on the market.
* Marginal Effect: Each additional day spent in pending status added approximately one extra day to the total adom.

1. sqft (Square Footage of the Home)

* Effect: Larger homes tended to stay on the market longer.
* Marginal Effect: For every 1,000 additional square feet, homes remained on the market for several extra days, indicating that larger homes may take longer to sell due to fewer potential buyers or higher prices.

Appendix

Individual analysis: ‘pricesold’

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

Individual analysis: ‘adom’

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.