Real Estate

Jacob Eliason and Gabriel Adams

April 2019

Executive Summary

Through use of a multiple linear regression model that accounts for heteroskedasticity and spatial correlation, we were able to successfully analyze our data and address the research questions posed in the the introduction. First, we determined that by using the characteristics included in our model roughly 93% of the variation can be explained. Next, we determined that above-ground square footage, the year a home was built, the presence of AC, and the number of cars the garage can hold all have positive effects on sale price. That is, holding all else constant, a home with AC is probably going to more expensive than one without and a newer home will be more expensive than an older. Next we determined that sale price variability does increase as living area increases. Finally, we used our model to make appropriate predictions for sale price at locations for which we did not have it.

Introduction and Problem Background

Purchasing a home is one of the most important financial decisions individuals and families make. For these purchases, determining the true value of property is vital. In this analysis, we are interested in appraising the value of various homes located in Ames, Iowa.

Our data consists of 465 observations. For each observation, values of the variables in the following table were taken.

Name	Description	Name	Description
Price	Sale price of home	Central.Air	Includes central air?
Lon	Longitude (transformed)	Full.Bath	Num. of full bathrooms above ground
Lat	Latitude (transformed)	Half.Bath	Num. of half bathrooms above ground
Gr.Liv.Area	Above ground living area in sq. feet	Bedroom.AbvGr	Num. of bedrooms above ground
House.style	1 Story, 2 Stories, Split Level	Kitchen.AbvGr	Num. of kitchens above ground
Year.Remod.Add	Remodel/original construction date	Garage.Cars	Size of garage (num. of cars)

In addition to the 465 complete observations mentioned previously, there are 52 additional observations with values for all of the above variables except for price. In this analysis, we will predict the appraised prices of those 52 homes using information from the other variables and our statistical model.

Before continuing with our formal analysis, it is appropriate to perform exploratory data analysis. Below is a plot of the houses in their respective locations. The color of each dot represents the price of that house, a gray dot means it is one of the 52 homes that are missing prices.

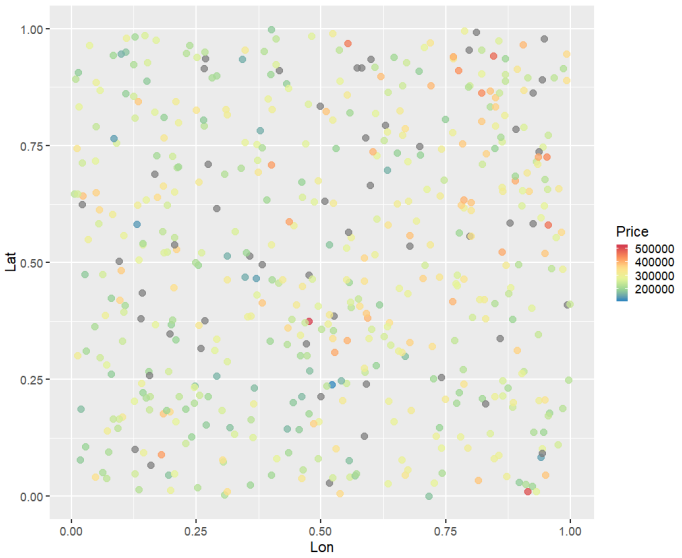

Figure 1: Home Prices

We can see that there are definitely pockets of higher prices, such as near the northeast corner, as well as areas with lower prices, such as in the south (in the middle). Our model will have to take these spatial patterns into account in order to be valid.

To clarify the existence of spatial correlation in our data, we fit a linear regression model without taking into account spatial correlation. We then plotted the raw residuals, which are shown below. We can clearly see patterns of residuals. For example, in the northeast corner the residuals are largely positive, while, in the southwest corner, the residuals are mostly negative. Spatial correlation is present and will need to be built into our model in order for our model to be valid. If spatial correlation is not taken into account, the independence assumption will not be valid. Any inference using our model, including hypothesis tests and confidence intervals, will be wrong. If spatial correlation is properly taken into account, those problems will go away. Accounting for spatial correlation will also help improve predictive performance (bias, RPMSE, average prediction interval width, and coverage should all improve).

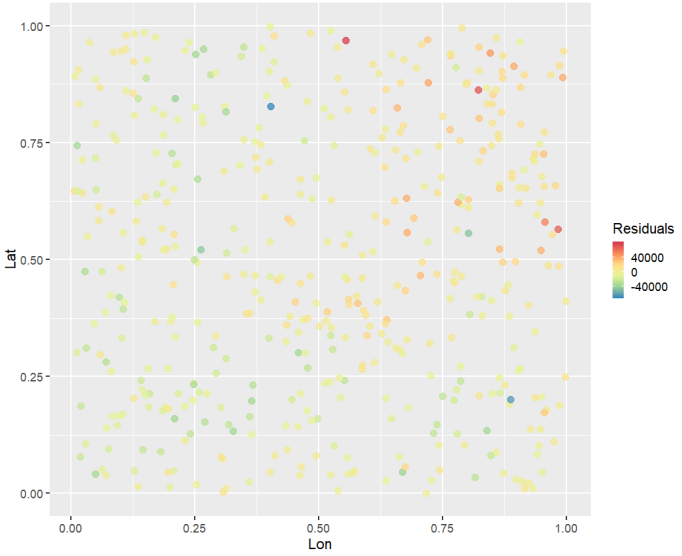

Figure 2: Residuals (Linear Model w/ No Correlation)

We can also look at a variogram. The plot below shows that variance is monotonically increasing as a function of distance, which is an indication of spatial correlation. In other words, home price is not independent of location, so our traditional assumption of independence is violated.

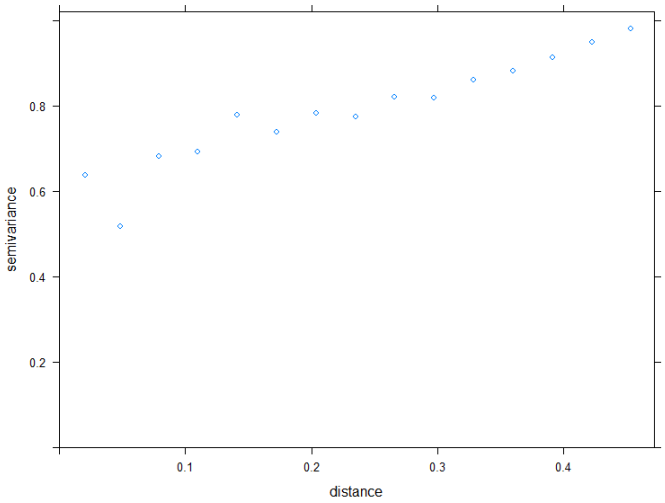

Figure 3: Variogram

Another characteristic of this data set is that the variability of sale price increases with the size of the home. We will show this formally later in the analysis. Thus, the assumption of equal variance will not be met without an appropriate modification to our model. If no modification is made, we will not be able to calculate correct confidence intervals and perform valid hypothesis tests. This is because both of those incorporate standard errors, which will be wrong because of the heteroskedasticity present in our data. By properly accounting for heteroskedasticity, we will be able to accurately perform inference.

For your information, we transform the two bathroom variables into a single quantitative "bath" variable, with each full bathroom worth 1 and each half bathroom worth 0.5. We shouldn't lose any information by doing so and it will simplify our analysis later on.

We have four primary research goals in this analysis. First, we want to determine how well the previously mentioned characteristics explain sale price. Next, we aim to specifically quantify the effects of the explanatory variables on price. Third, we want to see if the variability of sale price increases with the size of the home. Finally, we want to predict the prices for the 52 homes for which there is not a sale price.

To address these research questions, we will use a multiple linear regression model that takes into account both the spatial correlation and heteroskedasticity present in our data. We will augment an ordinary multiple linear regression model by adjusting the covariance matrix of our model. These adjustments, explained in detail later in this analysis, will help our model be valid. Then we will be able to use our model to perform inference, make predictions, and answer the research questions posed in the previous paragraph.

Statistical Model

Our statistical model is as follows:

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\Sigma} = \mathbf{DRD}$$

where

$$\mathbf{D} = diag(d_{ii}), \quad d_{ii} = exp\{2(GrLivArea_i)\theta\}$$

and

$$\mathbf{R} = \sigma^2((1-\omega)\boldsymbol{C} + \omega\boldsymbol{I})$$

where

$$\boldsymbol{C}_{ij} = exp\left\{-\frac{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|}{\phi}\right\}$$

$\mathbf{Y}$ is a vector of the values the response variable (price) for each of our observations. $\mathbf{X}$ is a design matrix with information from the explanatory variables, with the first column being a column of 1's corresponding to the intercept. $\boldsymbol{\beta}$ is a vector of effects, one for the intercept, and one for each quantitative explanatory variable or non-baseline level of a categorical variable. The matrix $\boldsymbol{\Sigma}$, called the covariance matrix, provides the variance of the model. $\boldsymbol{\Sigma}$ can be decomposed into two parts: $\mathbf{D}$ and $\mathbf{R}$, as shown above.

$\mathbf{D}$ provides a measure of variance for individual data points and will provide a way of quantifying the heteroskedasticity in our data. $\mathbf{D}$ is diagonal matrix, with the diagonal elements defined by $d_{ii} = exp\{2(GrLivArea_i)\theta\}$. If $\theta > 0$, then the variance increases as living area increases. If $\theta < 0$, the variance decreases as living area decreases.

$\mathbf{R}$ is a correlation matrix that quantifies the correlation between the data points. This is necessary due to the spatial correlation present in our data. σ^2 is a measure of overall variance.

ω is called the nugget. It, together with $\boldsymbol{I}$, is a measure of variance at a single point. It would seem that two measurements taken at the same exact point would have the same value. However, due to sampling error and other potential issues, a nugget is necessary to quantify variation at a single point. Obviously, different houses will never be at the same location, but the nugget will help keep the model stable.

$\boldsymbol{C}$ is a matrix describing the spatial correlation in our model. The off-diagonal elements represent a measure of spatial correlation between spatial locations. We are using the exponential correlation structure (justified later). $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ represents the Euclidean distance (in Lon/Lat units) between two points. ϕ is called the range parameter. It is a measure of how correlation decays over distance. A larger value of ϕ means that correlation decays at a slower rate (the range of correlation basically increases).

In order for our model to be valid, we will need to justify linearity, independence, normality, and equal variance. These assumptions are similar to those of standard multiple linear regression models, but use some special procedures to account for correlation. These assumptions will be justified in the next section.

Model Validation

We proceed to checking the assumptions of our model. From the added-variable plot shown below, we can see that the assumption for linearity is met. There are no non-linear patterns visible.

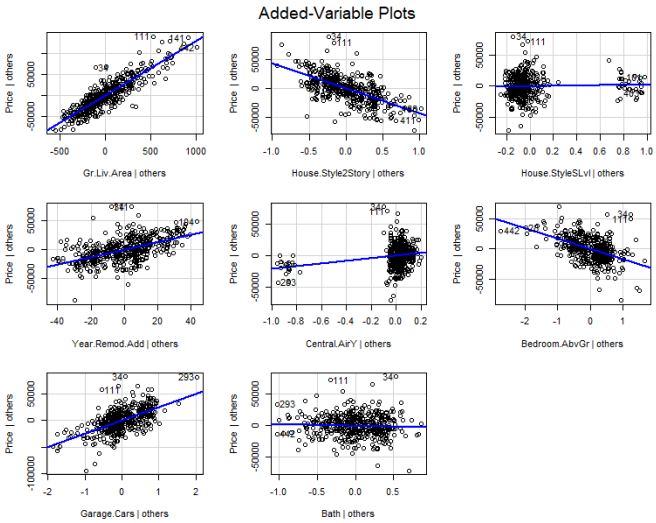

Figure 4: Added-variable plots

We extract the decorrelated residuals and plot them in space to see whether there is any lingering spatial correlation. We observe that there are no visible patterns or clustering in the residuals.

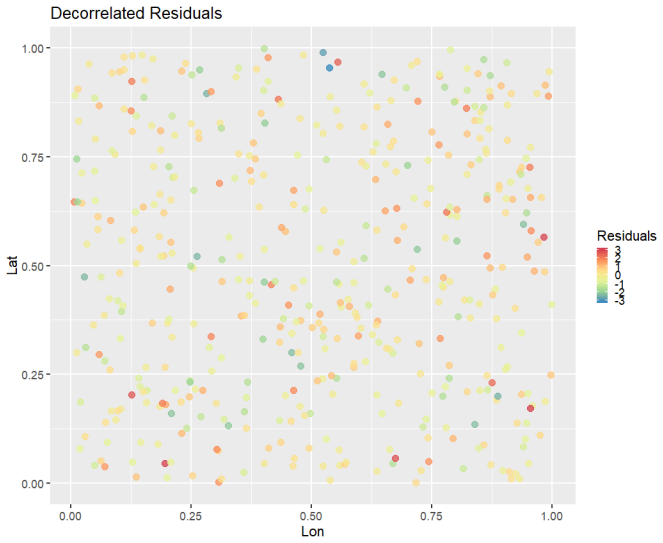

Figure 5: Plot of decorrelated residuals

We can also look at a variogram of the decorrelated residuals. There appears to be a flat trend, which indicates that variance is not increasing as a function of distance. Thus, we conclude that the spatial correlation is accounted for and our assumption of independence is met.

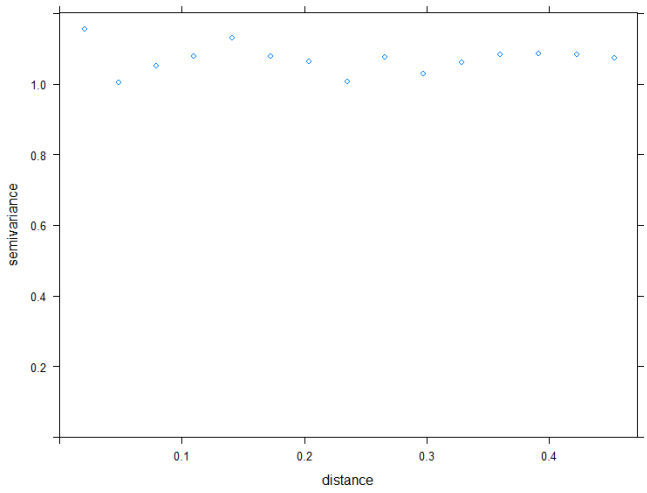

Figure 6: Variogram

A histogram of our decorrelated residuals looks approximately normal. Our assumption of normality in the residuals is met.

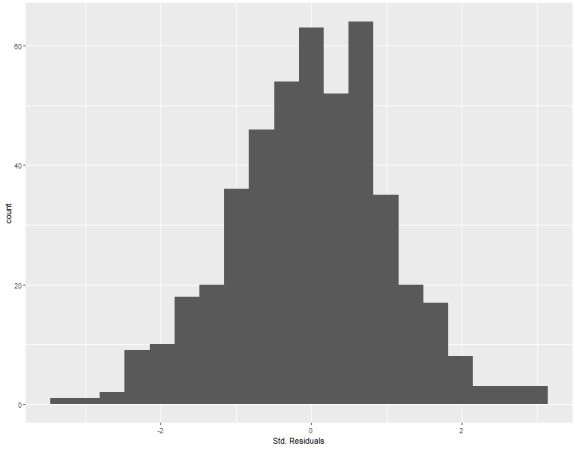

Figure 7: Histogram of decorrelated residuals

Finally, we plot the decorrelated residuals against the fitted values from our model and find that there is sufficient scatter as the fitted values increase. We conclude that our assumption of equal variance is met.

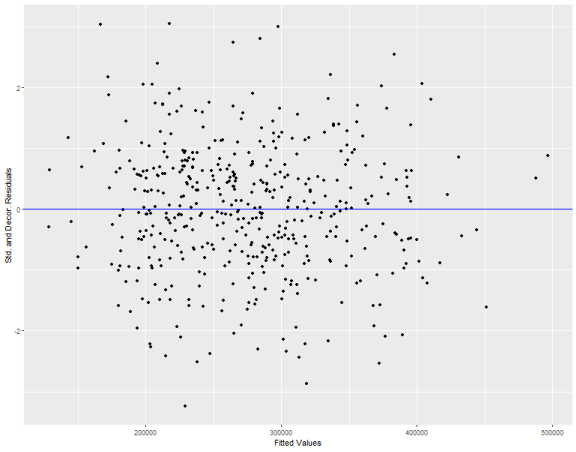

Figure 8: Fitted Values vs. Residuals

We find that our model is an excellent fit for this data set. We observe a pseudo-R-squared value of 0.932 using the following expression (squaring the correlation of predicted values and actual values):

$$cor(\hat{y}, y_i)^2$$

To further evaluate our model, we conduct a cross-validation study. We conduct 50 simulations using a 70-30 train-test split on our data to observe model performance compared to a naive linear model. Our results are displayed below. We can see clear improvement in predictive abilities across the board.

Measure	**GLS Model**	**Naive Model**
Bias	4.187	103.043
RPMSE	13537.022	17660.434
Coverage	0.959	0.275
Width	51574.327	11272.262

We find that our GLS model is relatively unbiased and has a lower root predictive mean squared error than the naive model. Most notable, 95% confidence intervals produced using the GLS model actually contained the observed values 95.9% of the time, compared to 27.5% of the time when using the naive model. Finally, we observe an average interval width of $51,574.

We also note that we compared two models that would account for heteroskedasticity and spatial correlation. We ended up using a model with a exponential correlation structure (as shown in the previous section) instead of a spherical correlation structure (similar model, only with a different correlation structure) because it minimized AIC. The AIC of the spherical model was 10074.79, while the AIC of the exponential model was 10071.65.

Our model assumptions appear to be justified. Heteroskedasticity and spatial correlation have been accounted for. Additionally, our model fits the data well and has good predictive capabilities. We are ready to continue our analysis and use our model to answer the research questions posed in the introduction.

Analysis Results

Before proceeding to our research questions, we report estimates and confidence intervals for various parameters of our model. These are included in the table below.

	Est	**Low**	**Up**
θ	0.0007	0.0006	0.0008
ϕ	0.256	0.100	0.652
ω	0.343	0.189	0.540
σ	5633.43	4195.91	7563.43

We are now ready to answer our research questions. The first question regards how well the home characteristics explain sale price. One way to answer this is to look at the pseudo-R-squared value we calculated earlier. That value is 0.932. This means that, roughly, 93% of the variation in sale price is explained by the variables in our model. This is pretty good.

The next question asks which factors increase the sale price of a home. To answer this question, we report 95% confidence intervals for our model coefficients below.

Coefficient	**Lower Bound**	**Upper Bound**
(Intercept)	-1425891.66	-1223758.25
Gr.Liv.Area	116.66	132.81
House.Style2Story	-45467.91	-39250.37
House.StyleSLvl	-3099.63	4312.18
Year.Remod.Add	662.04	766.10
Central.AirY	17368.86	25637.18
Bedroom.AbvGr	-17050.03	-13515.59
Garage.Cars	21021.22	24536.93
Bath	-4944.86	1162.26

As shown in the table, the home characteristics that correspond with increases in home price are above-ground square footage, the year the home was built or remodeled, whether it has central air conditioning, and how many cars the garage can hold. As a house increases in size by 1 square foot, we are 95% confident that the average price of homes will go up by between \$116.66 and \$132.81, with a best guess of \$124.73 (while holding all else constant). Similar interpretations exist for Garage.Cars (for each additional car) and Year.Remod.Add (as year increases by 1). We are 95% confident that a house with AC will be \$17,368.86 and \$25,637.18 more expensive than a house without it, with a best guess of \$21,503.01 (while holding all else constant). Other variables either have negative effects or are not statistically significant. Interestingly, the intercept is extremely negative, but this is largely offset by the cumulative effects of the Gr.Liv.Area and Year.Remod.Add variables.

We find that the variability of sale price increases as the size of the home increases. Our estimate for the variance parameter θ is 0.0007, with a 95% confidence interval from 0.0006 to 0.0008. Because this value is greater than zero, we know that home prices become more spread out at the positive extreme of the Gr.Liv.Area variable. The confidence interval does not contain zero, so it is statistically significant.

Finally, we are interested in prediction. Our data set is missing price information for 52 homes. We use our model to predict price for these missing values. Below are three plots showing the predicted prices, as well as the corresponding 95% prediction interval.

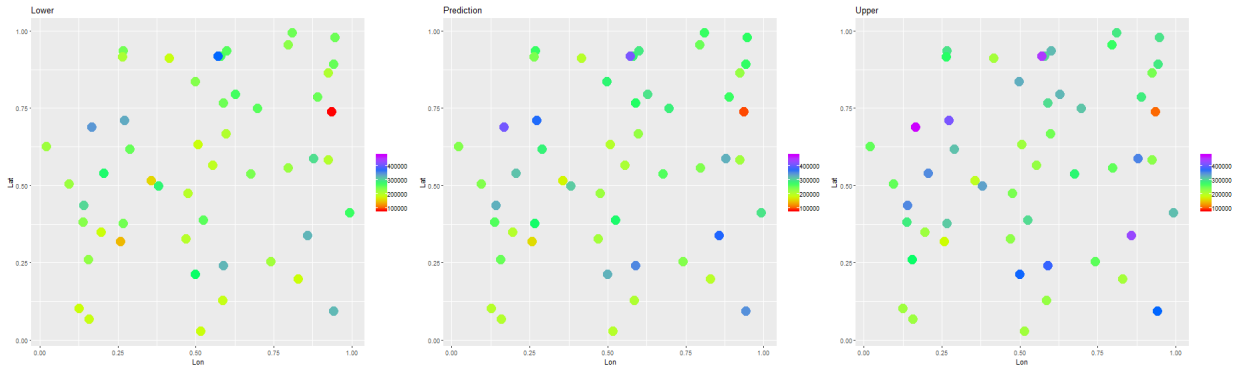

Figure 9: Predictions with Interval

Below we display our predicted results along with the existing observed values. Predicted values appear reasonable.

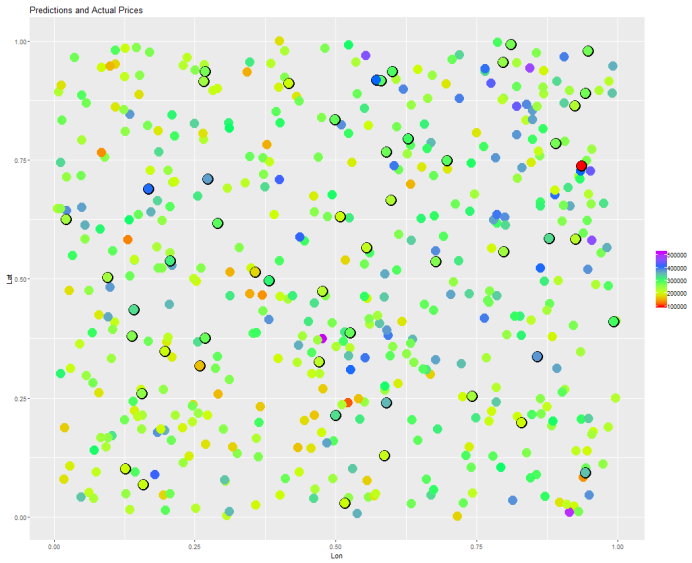

Figure 10: Predictions and Actual Prices

Conclusions

Through use of a multiple linear regression model that accounts for heteroskedasticity and spatial correlation, we were able to successfully analyze our data and address the research questions posed in the the introduction. First, we determined that by using the characteristics included in our model roughly 93% of the variation can be explained. Next, we determined that above-ground square footage, the year a home was built, the presence of AC, and the number of cars the garage can hold all have positive effects on sale price. That is, holding all else constant, a home with AC is probably going to more expensive than one without and a newer home will be more expensive than an older. Next we determined that sale price variability does increase as living area increases. Finally, we used our model to make appropriate predictions for sale price at locations for which we did not have it.

Future research could involve data from other locations, not just Ames, Iowa. Other variables could be collected such as total lot size. This additional information could greatly enhance what we've learned so far.