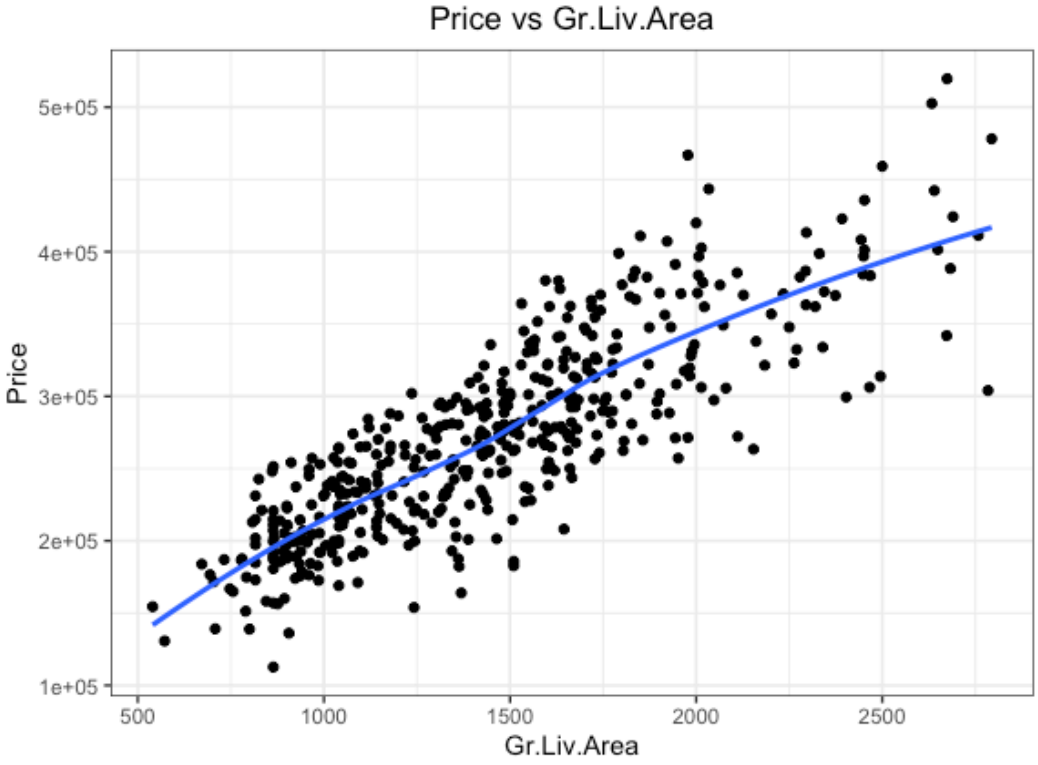
Michael Pixton
Stat 469

Analysis of Home Sale Prices

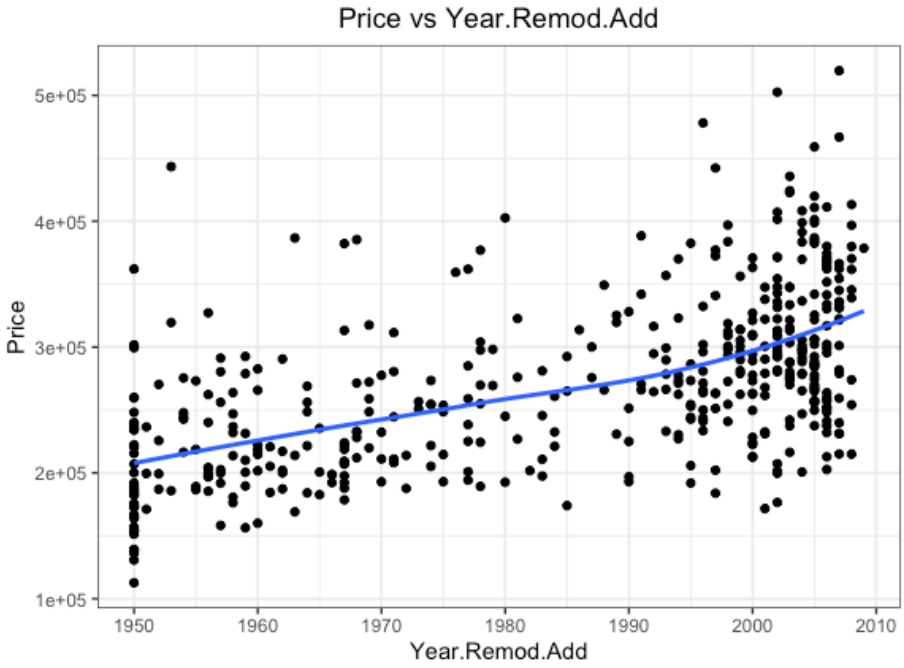
Section 1 – Introduction and Problem Background

Home appraisal is an important part of home buying and selling. It can help guarantee that the house is sold at the right price. Due to these needs. home appraisals are highly valued; we believe that by looking at some data taken in an area of home appraisals we may be able to design a model that will simulate the data give. This will help to answer the research questions concerning this dataset.

To begin, an exploratory data analysis was done, we large number of plots were created, the couple below were of particular note. Our dataset on have a couple of quantitative variables and the relationships with Price can be seen below.



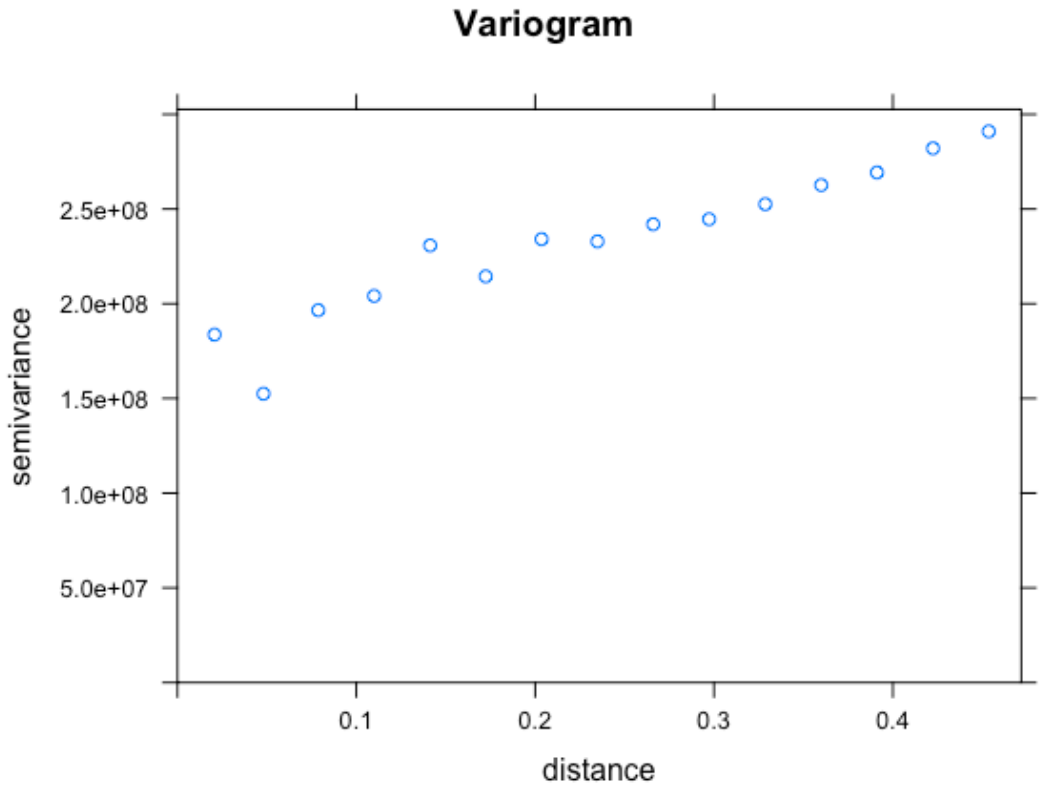
Michael Pixton
Stat 469



The Gr.Liv.Area (above ground living area) shows a strong linear correlation with Price and Year.Remod.Add (year the most recent remodel happened) shows a much less indication of correlations. Good thing to note is they are linear which is an assumption for our model. Due to the housing prices taken from an area, a type of map was created to show the price densities.

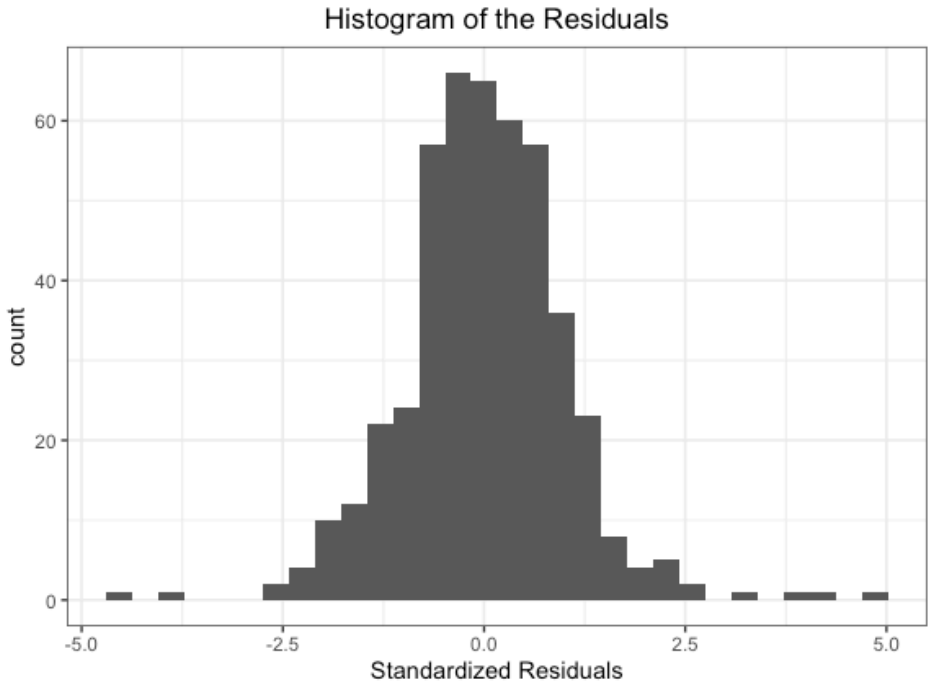
Michael Pixton
Stat 469

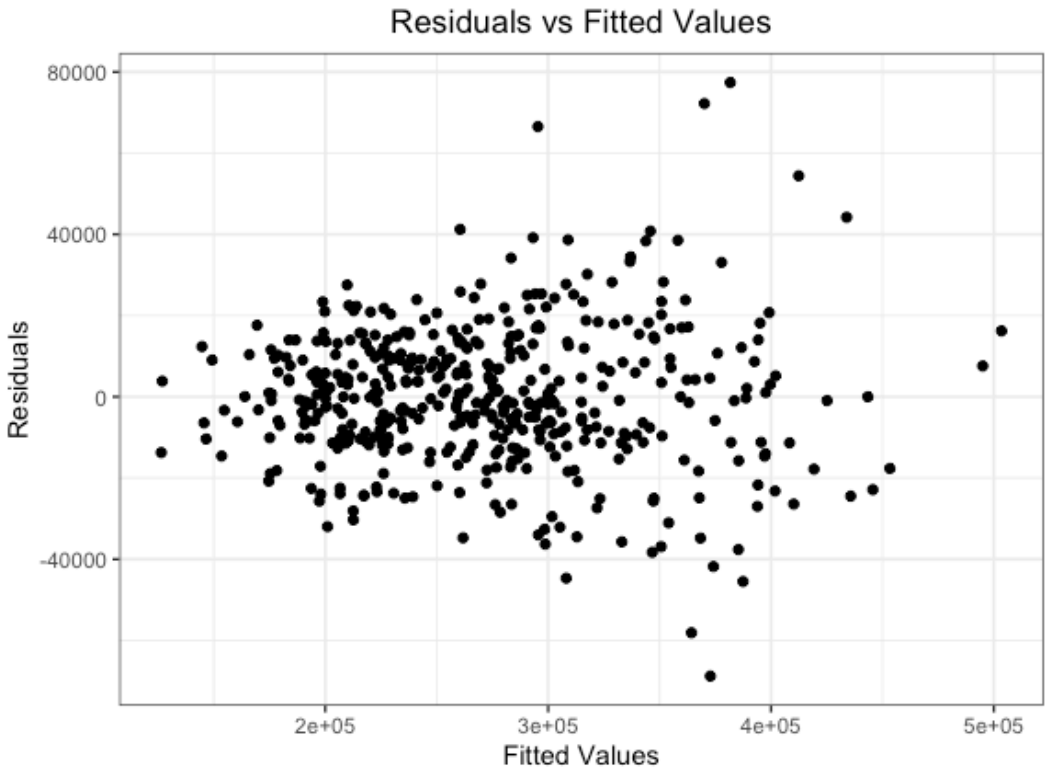
White on this map are locations that do not have data, but from what we can see, the data looks to have patches of darker and lighter areas. This is an effect of spatial correlation and must be accounted in our model. To double check this issue a variogram was created.



A straight horizontal line of dots would show there are no issues, but the slant that can be seem shows that spatial correlation must be taken into account, otherwise, our model's predictions and parameters will be off and lead to false assumptions. A simple model was fit to the data with the given explanatory variables, this model will help us check out assumptions. The two plots below are a histogram of the residuals and the residuals with their fitted values; these will help indicate issues with normality and equal variance.

Michael Pixton
Stat 469





Some serious issues can be seen in these two plots, one is that our histogram has points on the high and low end nearing 5 standard deviations, that is bit flat for our liking on the normal curve. However, the real problem is the Residuals vs. Fitted values that makes a clear

fan shape as you go to the left. This is a violation of the equal variance assumption and must be rectified with a change in our model to account for that fan shape.

In the end, we have decided that this data can be modeled to explain Price and attempt to answer the research questions. The model will be a Multiple Linear Regression (MLR) model that accounts for both the issues with heteroskedastic and spatial correlation.

Section 2 – Statistical Model

The model we wish to use will follow the form below:

$$\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{X\beta}, \boldsymbol{\sigma^2 DRD})$$

$$\mathcal{R}(\Phi, \omega) = \exp(- ||si - sj|| / \Phi)$$

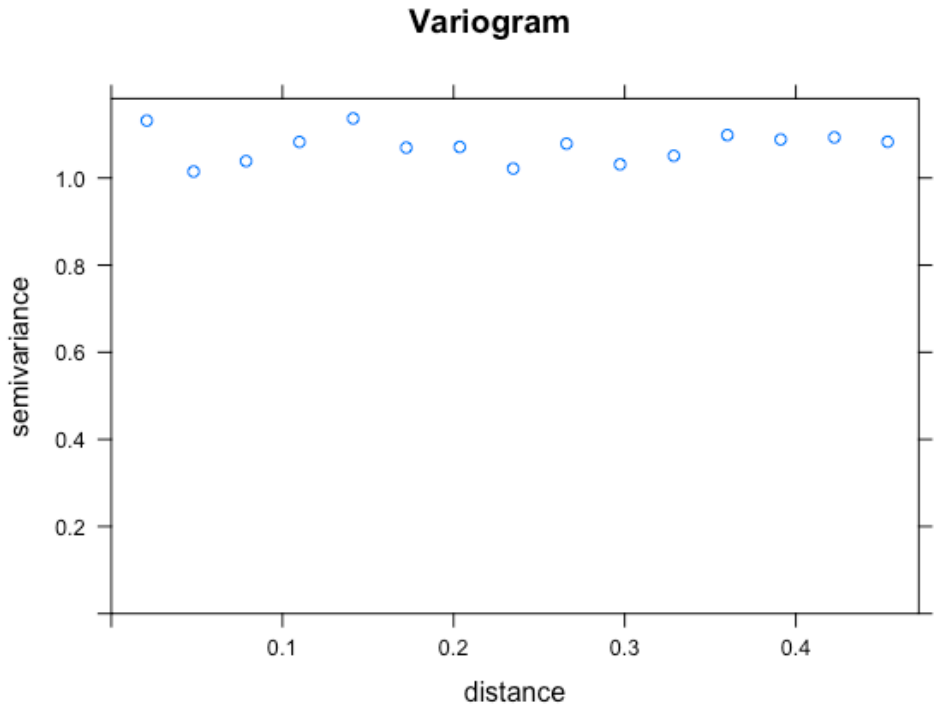
$$(D(\theta) = \exp\{2 * \log(x_{is})\theta\}$$

y = explanatory variable, Price
$\mathcal{N}$ = residuals should follow a normal distribution
$\mathcal{X}$ = The X matrix, with each independent variable from the dataset
β = The Beta matrix, the coefficient for each independent variable, it shows the effect of each variable in the X matrix on the y
σ^2 = Residual's variance is based upon a DRD structure
$\mathcal{R}$ = residuals following a correlations structure defined by R
Φ = parameter from the exponential correlation structure, each point is correlated by $p(s_i, s_j) = \exp(- ||s_i - s_j|| / \Phi)$
ω = parameter for the nugget in our correlation structure to insure we can have data from the identical point without perfect correlation where $Cor(s_i, s_j) = w$ if $|| s_i - s_j || = 0$, otherwise $(1-w)*p(s_i, s_j)$
$D(\theta)$ = the diagonal matrix, indicating the changes of variance as we move from each data point based upon a theta
$\exp\{2 * \log(x_{is})\theta\}$ = our exponential equation based upon several explanatory variables, that represents the change in variance of the residuals

When fitting this model, we took into consideration the validity of having all the variables in the model, a few tests were done and all variables were found to be either significant or nearly significant so they were all use when creating thing model. We also considered various structures for both Correlations matrix R and the Diagonal matrix D. The best correlations matrix was found to be Exponential and the Diagonal matrix had no other options fitting the data exponential was also chosen. In order to use this model, we must justify the LINE parameters: Linearity, Independence, Normality, Equal Variance. This will be proven in the next sections.

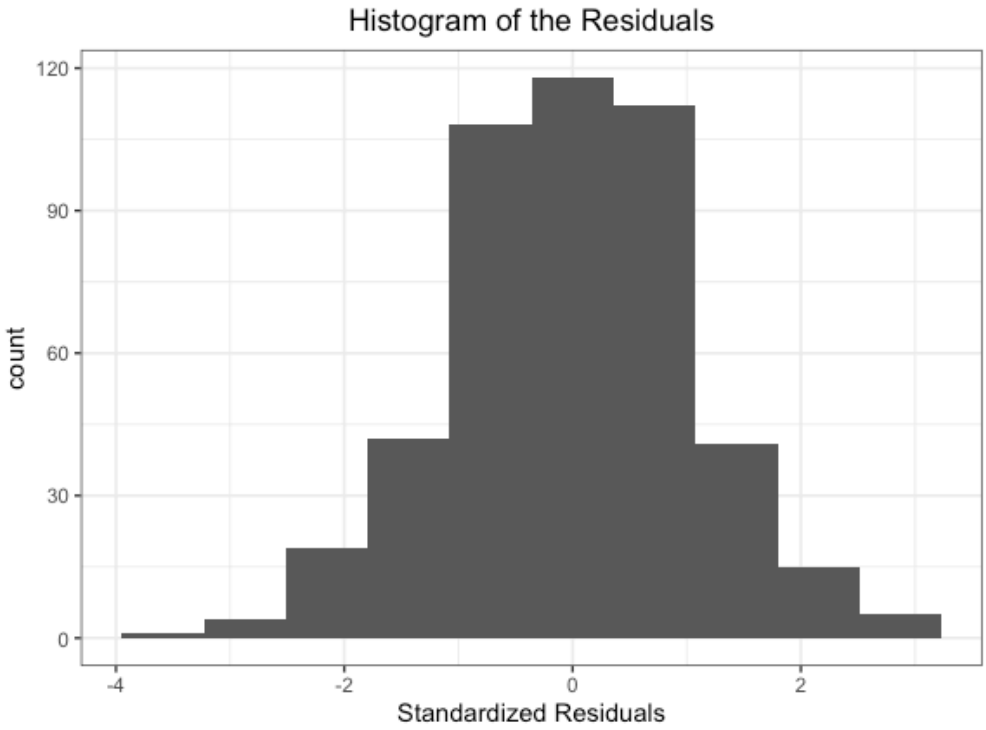
Section 3 – Model Validation

The assumptions must be proven meet. Linearity is first, and we can reference the plots given in the exploratory analysis. Linearity means that Price shows a linear relationship with the given independent quantitative variables; and the previous plots show that this assumption is met. Next is independence, meaning that each data point taken was independent of all other data points. Due to spatial correlation this was impossible, but our model attempted to reconcile this issue and a new variogram was created from our new model.



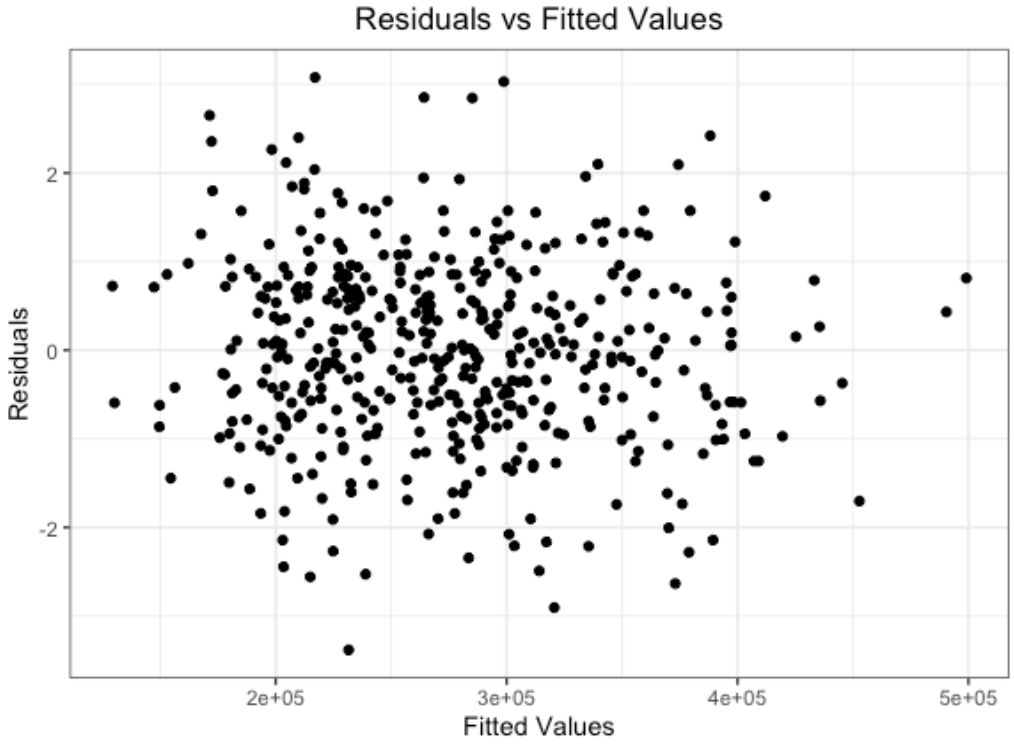
Michael Pixton
Stat 469

This plot shows that the, once slanted line, now looks pretty straight on the horizontal, meaning that our spatial correlations has been dealt with and we can accept that the independence assumption issue has is a nonissue. Next, normality can be quickly checked with a new histogram of the residuals.



This histogram looks a little better than the last one, still an outlier or two of the left, but not out to 5 standard deviations. This has gotten better due to our model fitting the data better and we can happily accept that the normality assumption has been met. Finally, equal variance is the last assumption we will look at. Equal variance was an issue before with Fitted values vs Residuals plot showing a fan like look; the new model's plot is shown below.

Michael Pixton
Stat 469



This plot looks significantly better than the previous one and shows that the residuals, once having a much greater spread on the right are now pretty equal across all the data. We believe that his has made out model have a much better fit and the equal variance assumption has been met.

In order to analyze the fit of our model, a cross validation series was run to test the model's fit with itself. We report a coverage of .93, much better than if the basic model had been used with reported a coverage of .76. We expect coverage to get 95% so 93% means that we are almost right where we expect. The bias is 5400, meaning that we are predicting off by about 5400 on a housed price. Considering the range is 113000, 52000, this is very good at pinpointing the true price within that large range.

R square is a measure of how well the explanatory variables explain Price. The basic model has an R-square of .93, meaning that the base model accounted for 93% of the variation of Price. As we have just proven, our model is better than the base, so we know that we have an R-square that is as good as .93, meaning we believe our model is very good at explaining housing prices.

Michael Pixton
Stat 469

Section 4 – Analysis Results

How well do home characteristics explain Housing Price? Well, as just stated, based upon the R-square of the base model we believe our model can explain at least 93% of the variation in housing prices. It could be quite close to 100 but we don't know for certain where our model lies between 93% and 100%.

What factors increase the sale price of a home? The chart below shows the various beta coefficients with 95% confidence intervals, indicating the effect of each of the explanatory variables have on Price.

Beta Hat Coefficients			
	lower	est.	upper
(Intercept)	-1421589	-1314214	-1206839
Gr.Liv.Area	116	124	132
Year.Remod.Add	654	708	761
Full.Bath1	-22395	5003	32401
Full.Bath2	-24990	2350	29690
Full.Bath3	-32552	2537	37626
Half.Bath1	-2204	802	3807
Half.Bath2	-45471	-20431	4610
Garage.Cars1	13522	18184	22846
Garage.Cars2	37089	41921	46753
Garage.Cars3	60433	67223	74014
Garage.Cars4	86607	119381	152155
House.Style2Story	-46045	-42634	-39224
House.StyleSLvl	-2607	1155	4918
Central.AirY	17524	21657	25789
Bedroom.AbvGr	-17004	-15153	-13302

We can interpret these intervals as, for example, as Gr.Liv.Area increases by once, we are 95% confidence the true change in Price will be between 116 and 132 dollars.

Does the variability of sale price increase with the size of the home (as given by living area)? The theta coefficient will tell us the effect on sale price with the size of the home. Theta was reported with a 95% confidence interval between 0.000604 and 0.000807, the whole range

is positive so we know that sale price sees significant variability as sale price with the size of the home.

What is your predicted/appraised sale price for the homes in the dataset that do not have a sale price? Predicted the homes that did not have a sale price from the dataset and got the following Prices with 95% confidence intervals attached as the lwr and upr categories, house numbers ordered as given in the dataset.

New Homes Prices					
House Number	**Lon**	**Lat**	**Predictions Price**	**lwr**	**upr**
1	0.925	0.863	227093.05	208245.55	245940.5
2	0.167	0.689	412895.61	350674.56	475116.7
3	0.586	0.128	214593.27	194153.85	235032.7
4	0.139	0.38	266252.19	240312.33	292192.1
5	0.268	0.936	271910.86	248103.45	295718.3
6	0.142	0.435	331655.38	306112.84	357197.9
7	0.022	0.625	241524.09	222747.65	260300.5
8	0.29	0.616	287554.95	256188.04	318921.9
9	0.83	0.198	199249.62	183270.44	215228.8
10	0.994	0.41	298676.76	274548.42	322805.1
11	0.6	0.936	293548.69	264499.18	322598.2
12	0.268	0.376	281633.31	250124.53	313142.1
13	0.858	0.337	380529.49	328470.94	432588
14	0.742	0.254	239594.95	219193.19	259996.7
15	0.382	0.497	310307.78	283594.77	337020.8
16	0.516	0.029	201885.22	184631.9	219138.5
17	0.526	0.386	281654.68	261017.31	302292
18	0.259	0.317	163108.68	144344.88	181872.5
19	0.59	0.766	278537.46	254559.18	302515.8
20	0.416	0.911	200877.73	185410.04	216345.4
21	0.266	0.915	244504.27	213251.29	275757.2
22	0.555	0.565	212307.2	195955.13	228659.3
23	0.796	0.955	253479.85	233979.94	272979.8
24	0.891	0.785	274621.57	255833.8	293409.3
25	0.677	0.536	264644.16	244937.04	284351.3
26	0.798	0.557	242478.48	225712.09	259244.9
27	0.629	0.794	298291.98	270572.95	326011
28	0.811	0.992	273462.49	253159.98	293765

29	0.599	0.666	226182.8	203576.02	248789.6
30	0.943	0.093	350279.1	323782.41	376775.8
31	0.943	0.891	275108.63	256649.32	293567.9
32	0.59	0.24	356918.7	327297.35	386540.1
33	0.879	0.585	329159.11	300567.74	357750.5
34	0.508	0.631	203013.08	185021.78	221004.4
35	0.159	0.068	204896.42	187462.79	222330.1
36	0.5	0.213	329877.6	282079.59	377675.6
37	0.206	0.538	317310.58	282172.27	352448.9
38	0.357	0.514	176530.08	159076.92	193983.2
39	0.476	0.474	222135.74	199840.58	244430.9
40	0.947	0.979	276809.31	254279.73	299338.9
41	0.698	0.749	289780.39	262781.38	316779.4
42	0.925	0.583	220966.18	205602.01	236330.4
43	0.499	0.835	282955.97	239211.53	326700.4
44	0.936	0.738	99098.57	87772.43	110424.7
45	0.096	0.503	240069.33	222485.27	257653.4
46	0.47	0.326	217612.54	201718.66	233506.4
47	0.581	0.917	288211.09	268896.29	307525.9
48	0.127	0.101	206479	187431.47	225526.5
49	0.572	0.917	414526.71	378299.35	450754.1
50	0.273	0.71	372746.27	333045.81	412446.7
51	0.156	0.26	255593.01	230630.4	280555.6
52	0.197	0.348	204176.06	185532.42	222819.7

We can interpret these as, for house #1, we are 95% confident the true price of the house is between $208000 and $246000.

Section 5 – Conclusions

Based upon the finding in this analysis, we have found that housing prices can largely be found dependent on the variables measured within this analysis. Having more bathrooms doesn't make your house's price much better, but having a 4-car garage is almost a guarantee that your housing price will be high. We also found that we could perform predictions on housed based upon the variables given with a good degree of accuracy.

Michael Pixton
Stat 469

Some follow ups to this analysis could be looking into different areas and comparing the results, particularly looking into downtown cities vs rural areas. These could help make the model stronger when comparing to these different area types.