Modeling Home Appraisals in Ames, IA

Carly Lundgreen and Logan Perry

Executive Summary

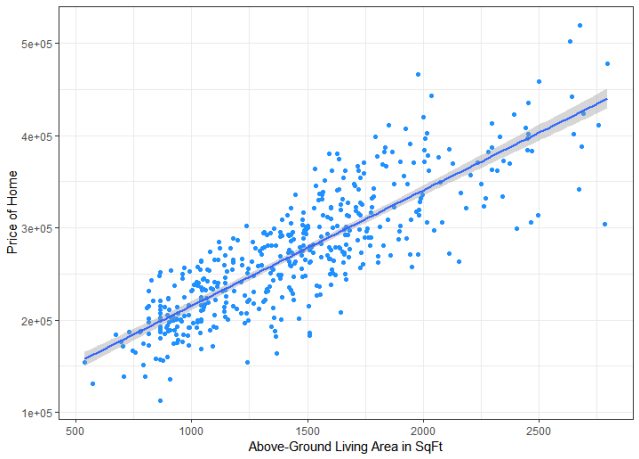
This analysis identifies significant factors that affect the sale price of a home using a relevant statistical model. Utilizing data gathered from the Ames, Iowa, area, we conclude that garage size, square footage of living area, whether the home has central air, and the year of any remodels are important factors that increase the sale price of a home. Larger homes tend to have more variable sale prices. Finally, using our model, we will predict sale prices for homes in our dataset that are missing appraisal values.

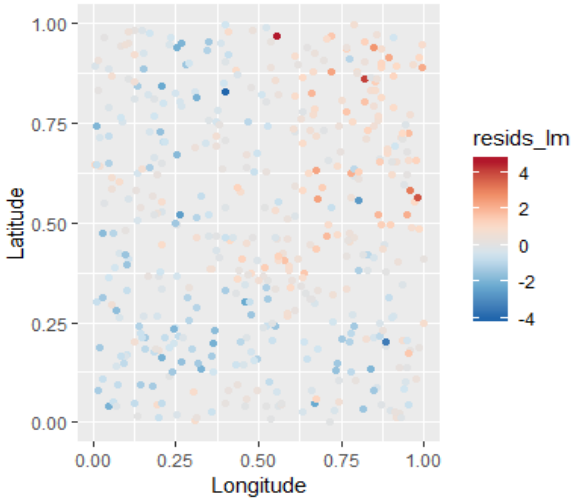
Introduction and Problem Background

A common occurrence when buying or selling a home is to have it appraised. A home appraisal is done by an unbiased third party so that an accurate value of the home can be given. Our goal is to build a model that can consistently appraise homes with accuracy. To do this, we have been given a data set with home sale prices (the response variable) and several characteristics of the home (explanatory variables). The variables and their descriptions are shown below. Note that the variable for the number of kitchens was not used in this analysis.

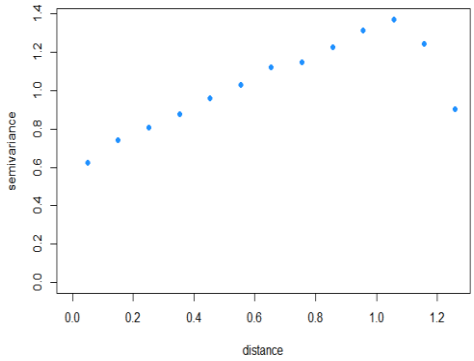
Variable Name	Description
Price	Sale price of the home
Lon	A transformed measure of longitude
Lat	A transformed measure of latitude
Gr.Liv.Area	Above ground living area in square feet
House.style	Style of dwelling (1 Story, 2 Stories, Split Level)
Year.Remod.Add	Remodel date (if no remodel than original construction date)
Central.Air	Include central air
Full.Bath	Number of full bathrooms above ground
Half.Bath	Number of full bathrooms above ground
Bedroom.AbvGr	Number of bedrooms above ground
Kitchen.AbvGr	Number of kitchens above ground
Garage.Cars	Size of garage in car capacity

There are two potential issues that we foresaw moving forward with the analysis: heteroskedasticity and spatial correlation. We noticed the heteroskedasticity when we made the scatterplot of price and above-ground living area, as shown below. This was confirmed when we ran a Breusch-Pagan test and calculated a p-value < 0.0001 meaning that we reject the null hypothesis of equal variance. We saw possible spatial correlation when we plotted the residuals from our multiple linear regression model and saw that almost all of the positive residuals were clumped in the top right of the plot (see scatterplot below).





We confirmed our suspicion by plotting a variogram which did not stably oscillate around a flat line (see plot to the right), providing evidence that there is spatial correlation. If we do not account for these two factors, our standard errors will be off. This means that while our point estimation will not change, our prediction and confidence intervals will be inaccurate because the correlation will affect the standard errors. To account for this correlation we will use a diagonal matrix **D** multiplied by the variance that will adjust for the heteroskedasticity. In addition to the **D** matrix, we will use an **R** matrix to account for the spatial correlation with a selected correlation function.



Statistical Model

We define our model as follows: $Y \sim N(X\beta, \sigma^2\Sigma)$, where Y is the nx1 vector of home appraisal values (i.e. our response variable). X is the n x (P + 1) matrix of explanatory variables. The matrix has a column of 1's, and is followed by the values of the P covariates that we utilize in the statistical model (style of home, garage capacity, living area, etc.). The vector β is defined as the (P+1) x 1 vector of effects that the covariates have on the sale price of homes in this area of Iowa. The parameter σ^2 is the variance of the homes' sale prices. Because we observed a small level of heteroskedasticity among these response variables, we must account for this unequal variance inside the covariance matrix of our model. The covariance matrix is defined as Σ = DRD, where D is the nxn diagonal matrix with elements d_{ii} on the diagonal and the zeros on the off-diagonal elements. This structure is displayed below. Each diagonal element, d_{ii}, is defined by an exponential variance function: $d_{ii} = \exp\{2x_i\theta\}$, where x_i is a covariate and θ indicates the way in which the response variable changes with x_i. For $\theta > 0$, the variance of y will increase with increases in x. We define the correlation matrix, R, as an nxn matrix where each

element $\rho(s_i, s_j)$ is the correlation between locations and i and j. The structure of this matrix is also shown below.

$$\mathbf{D} = \begin{pmatrix} d_{11} & 0 & 0 & \cdots & 0 \\ 0 & d_{22} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & d_{nn} \end{pmatrix} \quad \mathbf{R} = \begin{pmatrix} 1 & \rho(\mathbf{s}_1, \mathbf{s}_2) & \rho(\mathbf{s}_1, \mathbf{s}_3) & \dots & \rho(\mathbf{s}_1, \mathbf{s}_n) \\ \rho(\mathbf{s}_2, \mathbf{s}_1) & 1 & \rho(\mathbf{s}_2, \mathbf{s}_3) & \dots & \rho(\mathbf{s}_2, \mathbf{s}_n) \\ \vdots & & \ddots & \dots & \vdots \\ \rho(\mathbf{s}_n, \mathbf{s}_1) & \rho(\mathbf{s}_n, \mathbf{s}_2) & \rho(\mathbf{s}_n, \mathbf{s}_3) & \dots & 1 \end{pmatrix}$$

We have selected an exponential correlation structure between each location, which means that the elements of **R** are defined as follows:

$$\rho(\mathbf{s}_i, \mathbf{s}_j) = \exp\left\{ -\frac{\|\mathbf{s}_i - \mathbf{s}_j\|}{\phi} \right\}$$

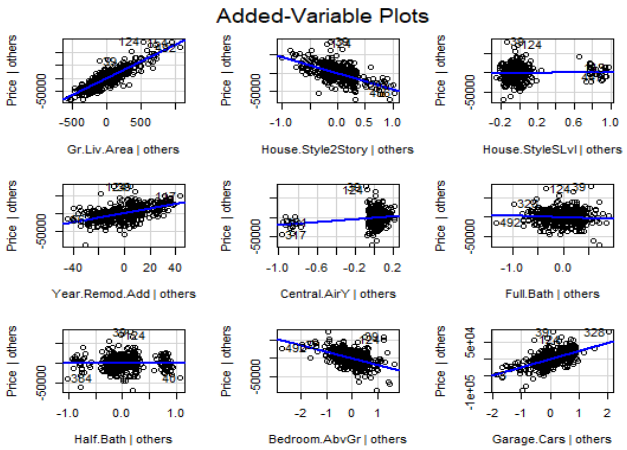
Where the range parameter ϕ indicates changes in spatial correlation. As ϕ increases, the range of the spatial correlation increases, and vice versa. It is important to note that this model assumes perfect correlation among measurements at same locations, differing only in the time at which the measurements were taken. The solution here is to add a variance nugget, which we define as follows:

$$\mathbb{C}\text{or}(\mathbf{s}_i, \mathbf{s}_j) = \begin{cases} \omega & \text{if } \|\mathbf{s}_i - \mathbf{s}_j\| = 0 \\ (1-\omega)\rho(\mathbf{s}_i, \mathbf{s}_j) & \text{otherwise} \end{cases}$$

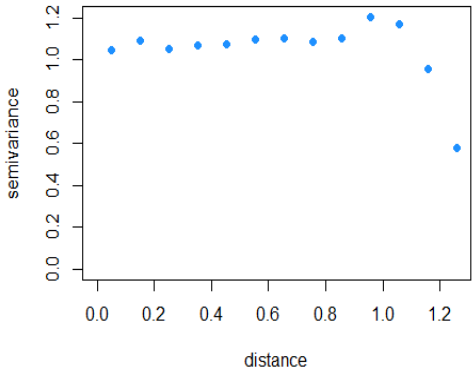
The parameter ω indicates the effect that same location variability, $|| s_i - s_j || = 0$, has on the correlation between housing prices at those locations. We chose the exponential spatial correlation structure because this model had the lowest AIC value, a method of comparing model fit, when compared to two other models that we fit with Gaussian and spherical correlation structures. This model assumes that the relationship between sale price and each covariate is linear and that the decorrelated residuals are normally distributed. We also must check that the issue of heteroskedasticity is eliminated so that the model conforms to the equal variance assumption of multiple linear regression models. We also will make sure that the model captures all the spatial correlation that is present among the residuals so that the assumption of independence, after decorrelation and standardization of the residuals, can be satisfied.

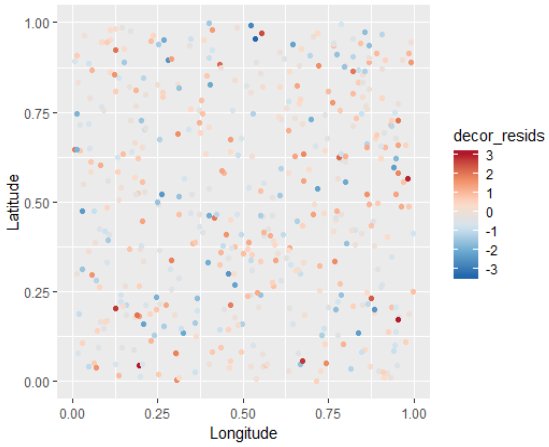
Model Validation

We check to make sure that the relationship between home sale price and the covariates is linear with the following added-variable plots. We can see that these relationships are sufficiently linear to proceed.

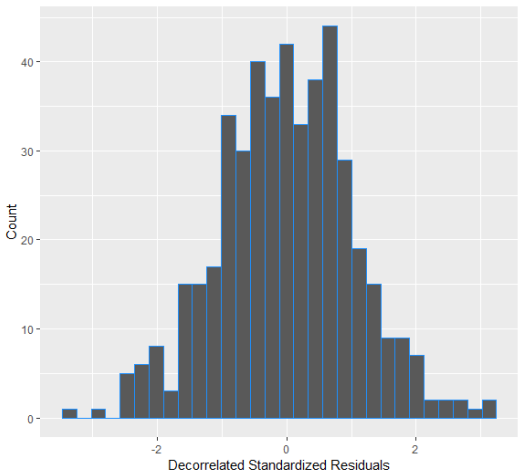


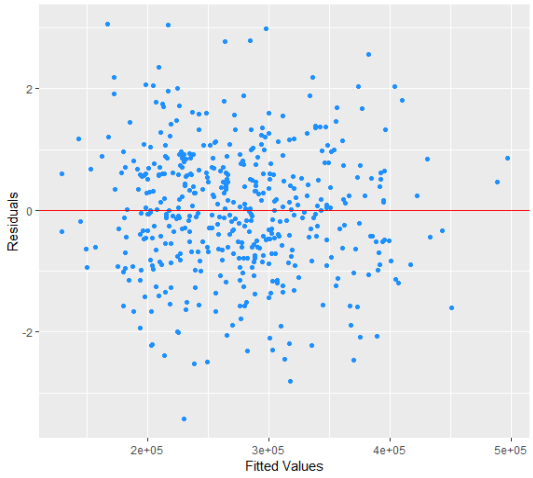
Independence between the sale prices of homes is not an assumption because we know that there is spatial correlation, but we account for this correlation in our model using a covariance matrix. We can see the correlation captured in our model in the following variogram compared to the one shown previously. Variance stays (generally) constant as a function of distance, except for the more extreme distances, but it appears that the spatial correlation has been taken care of. We double checked that the residuals had been sufficiently decorrelated by mapping them. Based on the map, seen below, we can see that values of each residual appear to no longer be affected by correlation based on spatial proximity.





The next assumption that we make with this model is normality of the decorrelated and standardized residuals. The following histogram of the residuals illustrates that the residuals are appear to be normally distributed. This is confirmed with a p-value of 0.9689 from a KS test for normality, where we fail to reject the hypothesis that the residuals are normally distributed. Finally, we must confirm that the problem of unequal variance has been accounted for with the diagonal matrix **D** in our model. After accounting for the unequal variance in our model, we created the following plot of fitted values vs. decorrelated standardized residuals. The residuals seem to be distributed evenly about the regression line, so we can confirm that the model has taken care of the unequal variance seen previously.





To assess how well our model fits the data, we can look at measures such as a pseudo R^2 which was calculated by squaring the correlation between the fitted values produced by our model and the observed prices. The pseudo R^2 was calculated to be 0.93 which shows an excellent model fit. We are also interested in making predictions, so we performed a cross validation on our model. We subset the dataset so that only rows with no missing values were used for the cross-validation study. The training set that was used for each iteration of the cross validation was 372 rows long, which is about 80% of the data. After performing 50 iterations, we calculated an average RPMSE of $13,777.96. This measure, essentially, summarizes how far off our predictions were on average. Compared to the range of the sale prices of homes, summarized in the table below, which is much wider than the calculated RPMSE, this means that our predictions were, comparatively, fairly accurate. One statistic to note was that the coverage for our prediction intervals was only 90% when we expected it to be 95%; this shows that our prediction intervals are not capturing the true value as often as we expect. However, the model is still performing quite well.

Min	1st Quartile	Median	3rd Quartile	Max
112684	220309	266018	313183	519677

Analysis Results

In order to determine how well each of the home characteristics explain sale price, we will simply discuss the pseudo-R^2 and predictive accuracy of our model. On their own, home characteristics using a normal linear model does a poor job of explaining home appraisal price in comparison to the GLS Model, which is our spatial heteroskedastic model. RPMSE, bias, and width are all smaller; the only place where the normal linear model does better is in coverage. Our model fits the data well, as we can see with the high pseudo-R^2.

	GLS_Model	MLR_Model
RPMSE	13777.96	18439.95
Bias	36.16	-183.58
Width	50937.31	69925.54
Coverage	0.90	0.94

We also wanted to identify which factors increase the sale price of a home. To answer this, we looked at the $\hat{\beta}$ coefficients with 95% confidence interval bounds. The results are displayed in the table to the right, and show that increasing above-ground living area, the remodel year, the number of cars you can fit in your garage, and having central air all significantly increase the price of a home.

	Lower Bound	Upper Bound	Fit	P-Value
(Intercept)	-1426988.83	-1224778.39	-1325883.61	0.00
Gr.Liv.Area	116.73	132.90	124.82	0.00
House.Style2Story	-46475.73	-39713.34	-43094.53	0.00
House.StyleSLvl	-2950.60	4484.62	767.01	0.69
Year.Remod.Add	662.89	767.01	714.95	0.00
Central.AirY	17189.23	25482.93	21336.08	0.00
Full.Bath	-5735.33	747.94	-2493.70	0.13
Half.Bath	-2493.84	3371.78	438.97	0.77
Bedroom.AbvGr	-17163.86	-13609.93	-15386.90	0.00
Garage.Cars	21119.65	24657.86	22888.76	0.00

We were also asked to identify whether the variability of sale price increases with the size of home, as given by living area. To answer this, we looked at the parameter θ, which explains the change in variance among the response variable. The estimate of θ for this analysis was 0.00073 with a 95% confidence interval of (0.00062, 0.00084). Because all elements within this interval are greater than 0, we can conclude that the variability in sale price increases with the size of the home. Because the effect of living area (our measure of home size) is positive, indicating that the sale price of the home will increase, on average, as living area within the home increases, we can conclude that the positive value of θ and associated confidence interval indicates that the variability of home sale price is larger for larger homes.

Lastly, we wanted to predict the appraisal prices for the homes in the dataset that had missing values. Here are the first few missing values that we predicted for. The full table values can be found right before the code appendix at the end of our report.

Price	Lon	Lat	Gr.Liv.Area	House.Style	Year.Remod.Add	Central.Air	Full.Bath	Half.Bath	Bedroom.AbvGr	Garage.Cars
227668.4	0.924844	0.863378	1144	1Story	1960	Y	1	0	3	1
411088.5	0.167304	0.688991	2855	2Story	2000	Y	2	1	4	3
210338.8	0.586383	0.128379	1114	1Story	2004	Y	1	1	3	0
265298.7	0.138623	0.380158	1576	SLvl	1961	Y	1	0	4	2
272742.3	0.267923	0.935788	1478	1Story	1992	Y	2	0	3	2
332278.3	0.141926	0.434625	1483	1Story	2001	Y	1	1	1	2

Conclusions

There are many factors that influence the appraisal of a home. Utilizing a heteroskedastic spatial regression model, we accounted for a heteroskedastic relationship between above-ground living area and sale price with an exponential variance function in the covariance matrix. Through maximum-likelihood estimation of the variance parameter from this function, we conclude that larger

homes are more variable in sale price. We accounted for spatial correlation among the residuals with an exponential correlation structure. This spatial correlation structure relied on transformed measures of longitude and latitude from the Ames, Iowa, area. We were able to capture this correlation and come up with a model that fit our data well. Finally, with this analysis we conclude that the most important factors that increase the sale price of a home are the amount of living area (in square feet), the year in which any remodeling occurred, the size of garage (in car capacity), and whether or not the home has central air. For further analysis, we recommend identifying more variables that may affect home appraisal prices and utilizing them in the statistical model. Expanding the analysis past the area of Ames, Iowa may also provide interesting results.

Table of Predictions for Missing Sale Prices

Price	Lon	Lat
227668.37	0.924844	0.863378
411088.45	0.167304	0.688991
210338.75	0.586383	0.128379
265298.71	0.138623	0.380158
272742.28	0.267923	0.935788
332278.32	0.141926	0.434625
241891.68	0.021804	0.624992
287803.99	0.290472	0.616261
200640.02	0.829704	0.197647
298402.65	0.993523	0.410240
293843.40	0.600428	0.935593
281702.41	0.267679	0.375898
381925.66	0.858123	0.337239
240012.39	0.741728	0.254045
312536.83	0.381857	0.496638
202124.17	0.516414	0.028564
281765.37	0.526078	0.386436
162426.25	0.259005	0.317224
278605.80	0.589652	0.766432
202227.19	0.416255	0.911168
243962.40	0.265583	0.914812
211808.23	0.555315	0.565006
254084.25	0.796344	0.955165
274347.00	0.890557	0.784675
264624.42	0.677441	0.536172
242320.13	0.797819	0.556543
298062.18	0.628504	0.794114
273433.12	0.810594	0.992328
227396.97	0.598581	0.665618
350429.45	0.942727	0.092632
274929.43	0.943416	0.890534
355642.63	0.590249	0.239839
329660.59	0.878809	0.585384
203354.75	0.507886	0.630834
205622.94	0.158764	0.067598
329270.41	0.499963	0.213068
318621.05	0.206451	0.538381
178007.75	0.357443	0.514040
222918.58	0.476307	0.473587
276946.33	0.947439	0.978660
289695.98	0.697629	0.748643
220202.85	0.925010	0.582830
284823.41	0.498617	0.834906
95883.57	0.936406	0.737570
241127.12	0.095587	0.503071
217142.83	0.470221	0.326172
288364.85	0.580848	0.916706
202793.45	0.126728	0.101070
413664.35	0.572124	0.916953
373924.40	0.273191	0.709950
256279.48	0.156463	0.259561
204003.32	0.196750	0.348235

Code Appendix

```
########## Libraries
library(ggplot2)
library(car)
library(nlme)
library(MASS) #stdres
library(dplyr)
library(geoR) #variog
library(lmtest)
library(kableExtra)
source("https://raw.githubusercontent.com/MJHeaton/glstools/master/predictgls.R")
source("https://raw.githubusercontent.com/MJHeaton/glstools/master/stdres.gls.R")

########## Research questions and how we will answer them
## 1 How well do the home characterisitics explain sale price? --> pseudo r^2
## 2 What factors increase the sale price of a home? --> beta coefficients
## 3 Does the variability of sale price increase with the size of the home (as given by living area)?
   # I think this has to do with identifying heteroskedasticity?
## 4 What is your predicted/appraised sale price for the homes in the dataset that do not have a sale price?
   # do cross validation and then prediction

########## Read in the data

housing <- read.csv(file = "https://mheaton.byu.edu/Courses/Stat469/Topics/3%20-%20SpatialCorrelation/3%20-
%20Project/Data/HousingPrices.csv",
                    header = TRUE)
kable(head(housing)) %>%
  kable_styling(bootstrap_options = c("striped"))

## we need to get rid of the NA's
housing_obs <- housing %>% filter(!is.na(Price))
housing_na <- housing %>% filter(is.na(Price))

#########
## EDA ##
#########

########## Plots
# not the best plot. remember that we can see correlation the best through the residuals
ggplot(data = housing_obs, mapping = aes(x = Lon, y = Lat, col = Price)) +
  geom_point() +
  scale_color_distiller(palette = "RdBu", na.value = NA)

# scatterplot of Longitude and Latitude colored in by the residuals
# notice that the positive residuals are all clumped together in the top right corner
housing_lm <- lm(Price ~. -Lon -Lat, data = housing)
resids_lm <- stdres(housing_lm)
ggplot(data = housing_obs, mapping = aes(x = Lon, y = Lat, col = resids_lm)) +
  geom_point() +
  labs(x = "Longitude", y = "Latitude") +
  scale_color_distiller(palette = "RdBu", na.value = NA) #possible spatial correlation

# box plot of price and whether or not they have central air
ggplot(mapping = aes(x = Central.Air, y = Price), data = housing_obs) +
  geom_boxplot(col = "orchid4") +
  labs(x = "Central Air", y = "Price")
```

```
# scatterplot of Price and Above-Gorund Living Area
ggplot(mapping = aes(x = Gr.Liv.Area, y = Price), data = housing_obs) +
  geom_point(col = "dodgerblue") +
  labs(x = "Above-Ground Living Area in SqFt", y = "Price of Home") #yikes possible heterskedasticity

# boxplot of house style and price
ggplot(mapping = aes(x = House.Style, y = Price), data = housing_obs) +
  geom_boxplot(col = "orchid4") +
  labs(x = "House Style", y = "Price")

# variogram to see the spatial correlation
coords <- housing_obs[,2:3]
variogram <- variog(coords = coords, data = resids_lm)
plot(variogram) #yikes look at that spatial correlation

########## Statistics
cor(housing_obs$Price, housing_obs$Gr.Liv.Area) #0.8372 strong positive linear relationship
cor(housing_obs$Price, housing_obs$Year.Remod.Add) #0.5658 moderately strong
cor(housing_obs$Price, housing_obs$Garage.Cars) #0.7613 strong positive
cor(housing_obs$Price, housing_obs$Bedroom.AbvGr) #0.2509 weak positive
cor(housing_obs$Price, housing_obs$Full.Bath) #0.6764 moderate positive

######################
## Fitting a Model ##
######################

## double check on the heterskedasticity that we saw earlier
bptest(housing_lm) #p-value < 0.0001 so we reject H0 that the variance is constant

## check the different correlation structures to find which one we should use

gls_exp <- gls(model = Price ~. -Lon -Lat, data = housing_obs, weights = varExp(form = ~Gr.Liv.Area),
               correlation = corExp(form = ~ Lon + Lat, nugget = TRUE), method = "ML")
gls_gauss <- gls(model = Price ~. -Lon -Lat, data = housing_obs, weights = varExp(form = ~Gr.Liv.Area),
               correlation = corGaus(form = ~ Lon + Lat, nugget = TRUE), method = "ML")
gls_spherical <- gls (model = Price ~. -Lon -Lat, data = housing_obs, weights = varExp(form = ~Gr.Liv.Area),
                      correlation = corSpher(form = ~ Lon + Lat, nugget = TRUE), method = "ML")

# find lowest AIC
AIC(gls_exp) #AIC  = 10072.45 <-- this is the winner
AIC(gls_gauss) #AIC = 10072.9
AIC(gls_spherical) #AIC = 10073.24

########################
## Model Assumptions ##
########################

## Lineartiy
avPlots(housing_lm) #what does it mean when there isn't really any line like for full and half baths?

## Independence
decor_resids <- stdres.gls(gls_exp)
variogram2 <- variog(coords = coords, data = decor_resids)
plot(variogram2, pch= 19, col = "dodgerblue") #much better except for the last point

## Normality of the residuals
qplot(x = decor_resids, geom = "histogram") #looks normal
ggplot(data = housing_obs, mapping = aes(x = decor_resids)) +
```

```
  geom_histogram(col = "dodgerblue") +
  labs(x = "Decorrelated Standardized Residuals", y = "Count")

## Equal Variance
# not really since we are adjusting for heteroskedasticity
ggplot(data = housing_obs, mapping = aes(x = fitted(gls_exp), y = decor_resids)) +
  geom_point(col = "dodgerblue") +
  geom_abline(slope = 0, intercept = 0, col = "red") +
  labs(x = "Fitted Values", y = "Residuals") #looks good

######################
## Cross Validation ##
######################

########## Figure out how much time this is going to take
system.time({
  gls_exp <- gls(model = Price ~. -Lon -Lat, data = housing_obs, weights = varExp(form = ~Gr.Liv.Area),
                 correlation = corExp(form = ~ Lon + Lat, nugget = TRUE), method = "ML")
}) #user 33.45, system 0.52, elapsed 34.23 --> we only care about elapsed time

########## Define variables
n_cv <- 50
n_samps <- round(x = nrow(housing_obs) * 0.8, digits = 0)
rpmse <- numeric()
bias <- numeric()
width <- numeric()
coverage <- numeric()
preds <- numeric
pb <- txtProgressBar(min = 0, max = n_cv, style = 3)

########## Cross Validate
set.seed(76)
for(i in 1:n_cv) {
  # split into train and test sets
  rows <- sample(x = nrow(housing_obs), size = n_samps)
  housing_train <- housing_obs[rows,]
  housing_test <- housing_obs[-rows,]

  # get the model
  my_gls_exp <- gls(model = Price ~. -Lon -Lat, data = housing_train, weights = varExp(form = ~Gr.Liv.Area),
                    correlation = corExp(form = ~ Lon + Lat, nugget = TRUE), method = "ML")

  # make predictions
  preds <- predictgls(my_gls_exp, newdframe = housing_test)

  # get info
  rpmse[i] <- (preds$Prediction - housing_test$Price)^2 %>% mean() %>% sqrt()

  bias[i] <- (preds$Prediction - housing_test$Price) %>% mean()

  width[i] <- (preds$upr - preds$lwr) %>% mean()

  coverage[i] <- mean((preds$upr > housing_test$Price) && (preds$lwr < housing_test$Price))

  # make a progress bar
  setTxtProgressBar(pb, i)
}
close(pb)

########## Cross Validation on just the lm()
```

```
rpmse_lm <- numeric()
bias_lm <- numeric()
width_lm <- numeric()
coverage_lm <- numeric()

set.seed(76)
for(i in 1:n_cv) {
  # split into train and test sets
  rows <- sample(x = nrow(housing_obs), size = n_samps)
  housing_train <- housing_obs[rows,]
  housing_test <- housing_obs[-rows,]

  # get the model
  my_lm <- lm(formula = Price ~. -Lon -Lat, data = housing_train)

  # make predictions
  preds_lm <- predict.lm(my_lm, newdata = housing_test, interval = "prediction")

  # get info
  rpmse_lm[i] <- (preds_lm[,'fit'] - housing_test$Price)^2 %>%
    mean() %>%
    sqrt()

  bias_lm[i] <- (preds_lm[,'fit'] - housing_test[,'Price']) %>% mean

  width_lm[i] <- mean(preds_lm[,'upr'] - preds_lm[,'lwr'])

  coverage_lm[i] <- mean((preds_lm[,'lwr'] < housing_test$Price) & (preds_lm[,'upr'] > housing_test$Price))
}

########## Check Prediciton Validations for gls()
mean(rpmse) #13777.96
mean(bias) #36.16
mean(width) #50937.31
mean(coverage) #0.9

########## Check Prediciton Validations for lm()
mean(rpmse_lm) #18439.95
mean(bias_lm) #-183.5839
mean(width_lm) #69925.54
mean(coverage_lm) #0.9369892

############################
## Statistical Inference ##
############################

########## How well do the home characterisitics explain sale price?
# pseudo r^2
(cor(housing_obs$Price, fitted(gls_exp)))^2 #0.93 which is pretty darn good

MLR_Model <- c(18439.95, -183.5839, 69925.54, 0.9369892)
GLS_Model <- c(13777.96, 36.16, 50937.31, 0.9)
comparison <- as.data.frame(cbind(GLS_Model, MLR_Model))
row.names(comparison) <- c("RPMSE", "Bias", "Width", "Coverage")
comparison <- round(comparison, digits = 2)
kable(comparison) %>%
  kable_styling(bootstrap_options = c("striped"))

########## What factors increase the sale price of a home?
# beta.hat coefficients
```

```
coef(gls_exp)
dt <- cbind(round(confint(gls_exp), digits = 2), round(coef(gls_exp), digits = 2))
dt <- cbind(dt, c(0.00, 0.00, 0.00, 0.69, 0.00, 0.00, 0.13, 0.77, 0.00, 0.00))
dt <- as.data.frame(dt)
names(dt) <- c("Lower Bound", "Upper Bound", "Fit", "P-Value")
kable(dt) %>%
  kable_styling(bootstrap_options = c("striped", "hover"))

########## Does the variability of sale price increase with the size of the home (as given by living area)?
# Yes, as seen by the plot below
ggplot(mapping = aes(x = Gr.Liv.Area, y = Price), data = housing_obs) +
  geom_point(col = "dodgerblue") +
  labs(x = "Above-Ground Living Area in SqFt", y = "Price of Home")
intervals(gls_exp)

########## What is your predicted/appraised sale price for the homes in the dataset that do not have a sale price?
preds_na <- predictgls(gls_exp, newdframe = housing_na)
housing_na$Price <- preds_na$Prediction
na_prices <- as.data.frame(preds_na$Price)
kable(housing_na) %>%
  kable_styling(bootstrap_options = c("striped")) #nice table of predictions

gls_na <- gls(model = Price ~. -Lon -Lat, data = housing_na, weights = varExp(form = ~Gr.Liv.Area),
              correlation = corExp(form = ~ Lon + Lat, nugget = TRUE), method = "ML")
resids_na <- stdres.gls(gls_na)
# I want to make a plot with both the decor_resids and the resids_na
ggplot(data = housing_obs, mapping = aes(x = Lon, y = Lat, col = decor_resids)) +
  geom_point() +
  labs(x = "Longitude", y = "Latitude") +
  scale_color_distiller(palette = "RdBu", na.value = NA) #look s a lot better
```