# Ames, Iowa Home Price Modeling

**Preeti Swaminathan & Patrick McDevitt**

**Homework 13**

## Analysis Question 1

## Restatement of Problem:

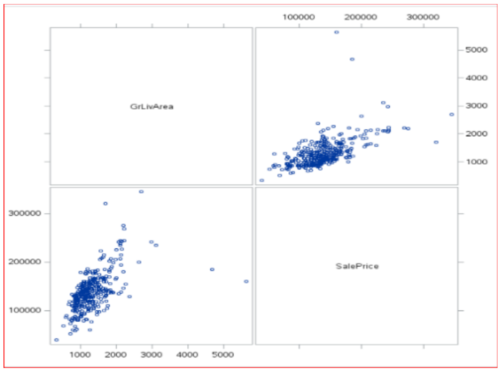
Century 21 in Ames Iowa wants to build a model to predict sales price of 3 neighbourhood (BrkSide, NAmes, EdWard) in Iowa based on living area. They have provided with historical data of sales done in these neighborhood so far.

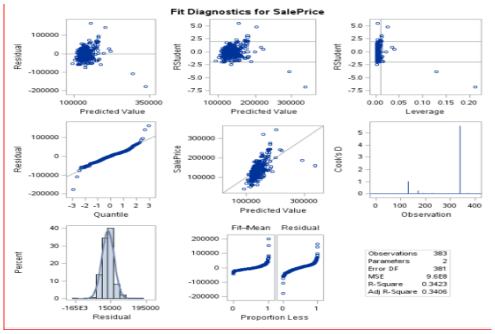
To-do: Build a model which uses (independent variables) living area in sq.ft and the neighborhood and predicts sales price (Dependent variable)

## Specify the model:

SalesPrice= $\beta 0$+$\beta 1$*GrLivArea+ $\beta 2$* neighborhood

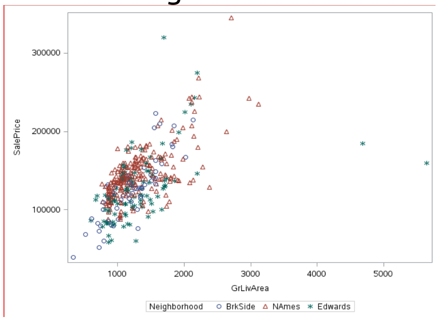
## Checking Assumptions

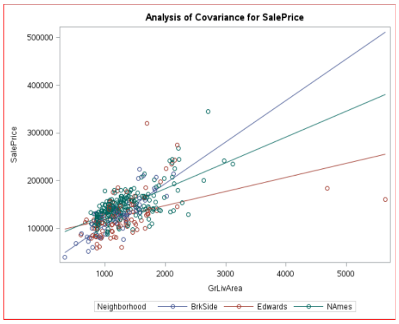



Fit Diagnostics for SalePrice

| | |
|---|---|
| Observations | 383 |
| Parameters | 2 |
| Error DF | 381 |
| MSE | 9.6E8 |
| R-Square | 0.3423 |
| Adj R-Square | 0.3406 |

1. From scatter plot, QQ plot and histogram, data is normally distributed
2. Interactions: looking at analysis of covariance plot, High leverage mild departures:
3. Looking at scatter plot b/w Sales price and living area, there are a few outliers, skewness, we can consider doing log transformation on sales price.
4. Looking at the plots, data is clustered in one group with outliers. This data needs to be log transformed




Analysis of Covariance for SalePrice

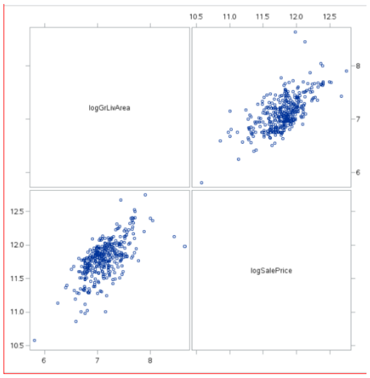
**Categorical variable analysis:**

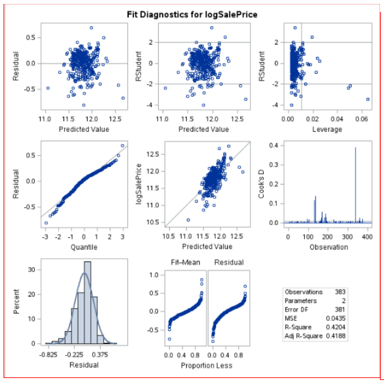
Data is concentrated around living area of < 30 sq foot and sales price for < 300k.

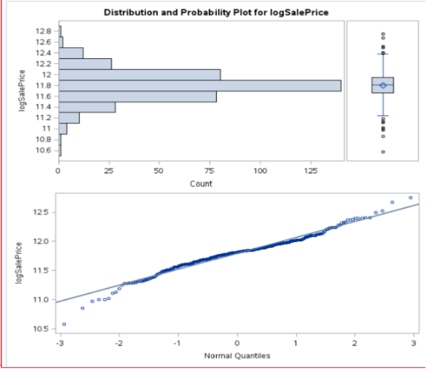
Edward neighborhood has a

few bigger houses
and higher price.

## Model 1: Log transformed on Sales price and living area.

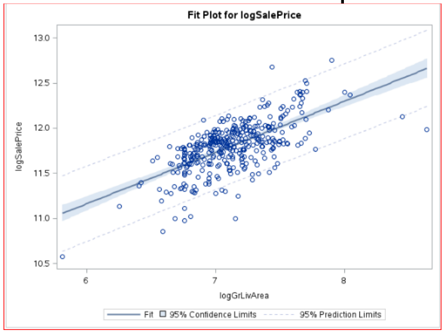


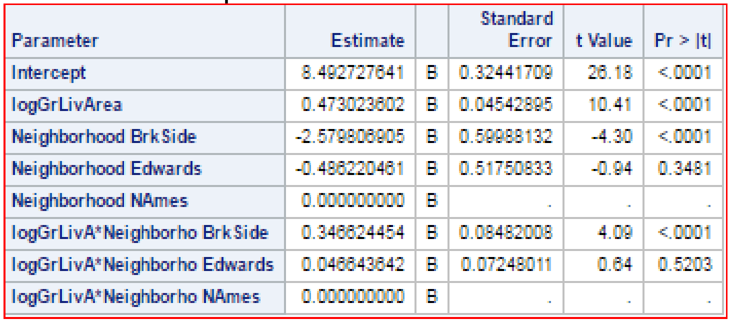




## Checking Assumptions

1. Normality: From QQ plot, scatter and histogram, we can assume normality. Although histogram shows a little skewness it's not very strong evidence against normality.
2. Linearity: Looking at the residual plots we can make an assumption of linearity.
3. Equal variance: After log transformation, since none of the plot looks too bad, we continue with our model.
4. CookDs shows couple of extrems. We will keep a check on it.

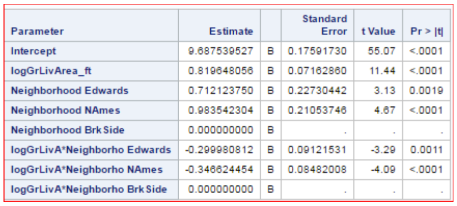




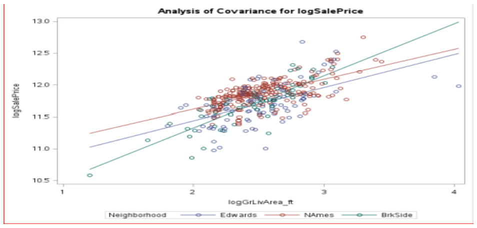
| Parameter | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 8.492727641 | B | 0.32441709 | 26.18 | <.0001 |
| logGrLivArea | 0.473023602 | B | 0.04542895 | 10.41 | <.0001 |
| Neighborhood BrkSide | -2.579806905 | B | 0.59988132 | -4.30 | <.0001 |
| Neighborhood Edwards | -0.486220461 | B | 0.51750833 | -0.94 | 0.3481 |
| Neighborhood NAmes | 0.000000000 | B | . | . | . |
| logGrLivA*Neighborho BrkSide | 0.346624454 | B | 0.08482008 | 4.09 | <.0001 |
| logGrLivA*Neighborho Edwards | 0.046643642 | B | 0.07248011 | 0.64 | 0.5203 |
| logGrLivA*Neighborho NAmes | 0.000000000 | B | . | . | . |

### T test and pValue

P value for Neighborhood BrkSide and logGrvLivA * Edwards is > 0.05.  We will make **BrkSide** as reference and recalculate the model



| Parameter | Estimate | | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 9.687539527 | B | 0.17591730 | 55.07 | <.0001 |
| logGrLivArea_ft | 0.819648056 | B | 0.07162860 | 11.44 | <.0001 |
| Neighborhood Edwards | 0.712123750 | B | 0.22730442 | 3.13 | 0.0019 |
| Neighborhood NAmes | 0.983542304 | B | 0.21053746 | 4.67 | <.0001 |
| Neighborhood BrkSide | 0.000000000 | B | . | . | . |
| logGrLivA*Neighborho Edwards | -0.299980812 | B | 0.09121531 | -3.29 | 0.0011 |
| logGrLivA*Neighborho NAmes | -0.346624454 | B | 0.08482008 | -4.09 | <.0001 |
| logGrLivA*Neighborho BrkSide | 0.000000000 | B | . | . | . |



Pvalue < 0.05 for all variables. Living Area and neighborhood are significant predictors of sales price.

## Fit the model

Log(SalesPrice)= $\beta0$+$\beta1$*Log(GrLivArea)+ $\beta2$* neighborhood BrkSide+ $\beta3$ neighborhood Edwards + $\beta4$ log(GrLivA) * neighborhood BrkSide + $\beta5$ log(GrLivA) * neighborhood Edwards

Log(SalesPrice) = 9.69 +  0.820*Log(GrLivArea)+ 0.712* neighborhood Edwards +0.984 neighborhood NAmes - 0.30 log(GrLivA) * neighborhood Edward- 0.347 log(GrLivA) * neighborhood NAmes

## Parameter Interpretation:

Log(SalesPrice) = 9.69 +  0.820*Log(GrLivArea)+ 0.712* neighborhood Edwards +0.984 neighborhood NAmes - 0.30 log(GrLivA) * neighborhood Edward- 0.347 log(GrLivA) * neighborhood NAmes

### Edward:

Log(salesPrice) = 9.69 + 0.820*Log(GrLivArea)+ 0.712 - 0.30 log(GrLivArea)
Log(salesPrice) =  10. 402 + 0.52 * Log(GrLivArea)
SalesPrice = e ^  10. 402 + 0.52 * Log(GrLivArea)
SalesPrice = e ^ 10.402 * grLivArea ^ 0.52

**SalesPrice {EdWard} = 32925 * grLivArea ^ 0.52**

For a 1sqr.ft house salesPrice of EdWArd = 32925$
Doubling it, GrLivingArea = 2, SalesPricec increases to 32926.43$

### Names

Log(SalesPrice) = 9.69 +  0.820*Log(GrLivArea) +0.984 - 0.347 log(GrLivArea)
Log(SalesPrice) = 10.674 + 0.473*Log(GrLivArea)
SalesPrice = e^10.674 * GrLivArea ^ 0.473

**SalesPrice{NAmes }  = 43217 * GrLivArea ^ 0.473**

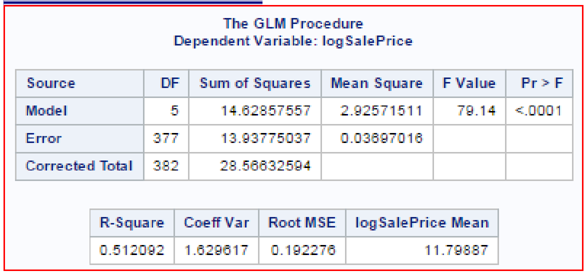
For a 1sqr.ft house salesPrice of EdWArd = 32925$
Doubling it, GrLivingArea = 2, SalesPricec increases to 32926.43$

### BrkSide

Log(SalesPrice) = 9.69 +  0.820*Log(GrLivArea)
SalesPrice = e^ 9.69 * GrLivArea ^ 0.82

**SalesPrice {BrkSide} = 16155 *  GrLivArea ^ 0.82**

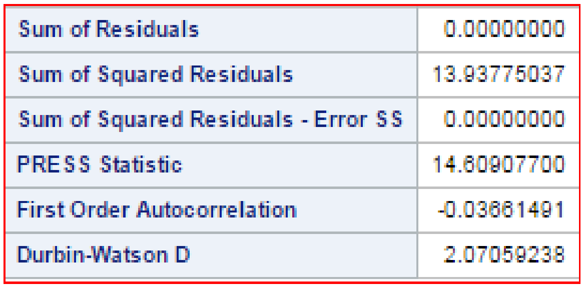
## R2 and Root MSE



The GLM Procedure
Dependent Variable: logSalePrice

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 14.62857557 | 2.92571511 | 79.14 | <.0001 |
| Error | 377 | 13.93775037 | 0.03697016 | | |
| Corrected Total | 382 | 28.56632594 | | | |

| R-Square | Coeff Var | Root MSE | logSalePrice Mean |
|---|---|---|---|
| 0.512092 | 1.629617 | 0.192276 | 11.79887 |



| | |
|---|---|
| Sum of Residuals | 0.00000000 |
| Sum of Squared Residuals | 13.93775037 |
| Sum of Squared Residuals - Error SS | 0.00000000 |
| PRESS Statistic | 14.60907700 |
| First Order Autocorrelation | -0.03661491 |
| Durbin-Watson D | 2.07059238 |

### Summary of anlysis:

1) R2 = 0.512, 51.2 % of variation in salesprice is affected by living area and neighborhood.
2) It appears that higher squart ft the sales price increases for all 3 neighborhood.
3) BrkSide neighborhood increases salesprice significantly compared to NAmes and

EdWard for this dataset.

4) Model without interactions was also developed. It did not produce reasonable results hence was discarded.