MSDS 6371 Project Report

Analysis of Home Sales in Ames, Iowa

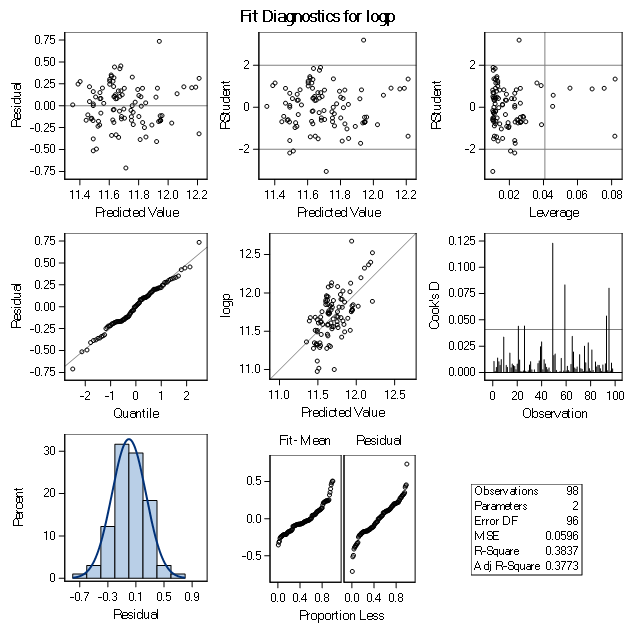
Richard Palmer

Jonathan Tan

Anthony Yeung

Analysis 1

#### Background

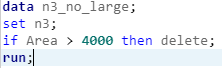
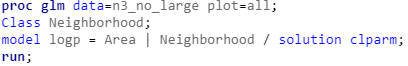
The goal of this first analysis was to determine the relationship between the sale price of a house and its square footage. Additional consideration was also given to the neighborhood each house was located in. There were a total of 381 observations within the three neighborhoods of NAmes, Edwards, and BrkSide. 

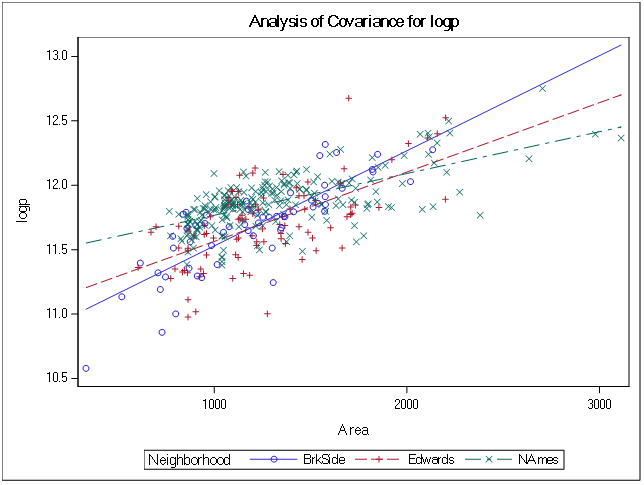
#### Checking Assumptions

* Linear relationship between variables is satisfied by graph of covariance for logp and area, which shows a roughly linear relationship
* Multivariate normality is satisfied by the qqplot of residuals
* Minimal collinearity
* Independent data must be assumed in order to proceed
* Equal variance   
  (maybe ggplots here

#### Model

LogSalePrice vs GrLivArea

Several houses had unusually large areas, and were removed as outliers  
  
  




#### 

Adj. R2 = 0.5344

Y is the final sale price of the house

𝛽0 is the y-intercept, or the theoretical value of a house with an area of zero square feet located in the NAmes neighborhood

𝛽1 is square footage of the house

𝛽2 is the coefficient representing the modifier to sale price if the house is located in BrkSide  
𝛽3 is the coefficient representing the modifier to sale price if the house is located in Edwards  
𝛽4 is the coefficient representing how the area of the house affects the sale price if the house is located in BrkSide

𝛽5 is the coefficient representing how the area of the house affects the sale price if the house is located in Edwards

Formula for each individual neighborhood:



#### Model Metrics (pg8 or pg 95?)

The independent variable, the area of the house, has been log transformed to meet normality assumptions of linear regression. Additional variables in the equation can be interpreted as meaning that houses in the BrkSide and Edwards neighborhoods have prices that are less affected by the square footage of the house. There are also separate modifiers for the interaction between the neighborhood the house is located in and the square footage of the house. We can see that the negative modifiers on the variable for BrkSide and Edwards indicate that the average sale price of houses there is lower than for NAmes.

Confidence intervals

Intercept - We can be 95% certain that the true mean of the intercept will fall between 11.443079 +/- 0.039466\*289.94 = (-0.3488, 23.247)

𝛽2 - We can be 95% certain that the true value of the logged area multiplier for BrkSide will fall between -0.577635 +/- 0.07762\*-4.97 = (-0.1918, -0.9634)

𝛽3 - We can be 95% certain that the true value of the logged area multiplier for Edwards will fall between -0.385474 +/-0.077622 \*-4.97 = (-0.0003, -0.7712)  
   
We can interpret this as every 1% increase in the area of the house will increase the sale price by 0.005($?), or that additional 100sq ft of living space, the price of any given house can be expected to increase by ???

#### Conclusion

Generally positive correlation across all neighborhoods, NAmes has the higher prices/higher correlation, followed by Edwards, then BrkSide) Given the data provided, the sale price of a house (in this area) generally positively correlates to the living area/square footage of the house. The final formula reached for the correlation between a house’s sale price, living area, and the neighborhood it is located in.

### Analysis 2

#### Background

The second item of analysis to build a predictive model for the sale prices of homes in the three neighborhoods NAmes, BrkSide, and Edwards based on several attributes of the houses in the provided dataset.   
 Models were constructed using forwards selection, backwards elimination, and stepwise regression.

#### Checking Assumptions

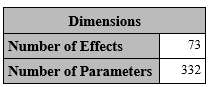
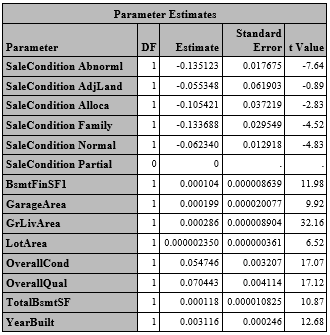
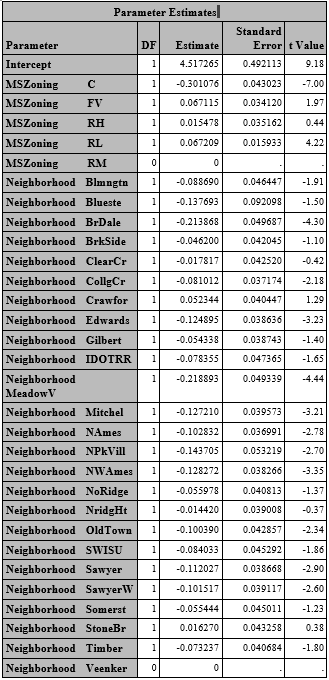
* Linear relationship is justified with scatterplot of the sale price with several of the primary attributes, such as area and number of rooms
* Multivariate normality
* Minimal collinearity
* Independent data
* Equal variance

#### Model the model used integrated the following attributed of each house in the training set:

Intercept BldgType BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath BsmtHalfBath BsmtQual CentralAir Electrical ExterCond ExterQual Foundation Heating HeatingQC LandContour LandSlope LotShape MasVnrType MSZoning Neighborhood SaleCondition SaleType Street BsmtFinSF1 BsmtFinSF2 BsmtUnfSF EnclosedPorch GarageArea GarageYrBlt GrLivArea LotArea LotFrontage LowQualFinSF MasVnrArea MiscVal MoSold OpenPorchSF OverallCond OverallQual PoolArea ScreenPorch TotRmsAbvGrd WoodDeckSF x1stFlrSF x3SsnPorch YearBuilt YearRemodAdd YrSold.   
  
The sale price of the houses was log transformed

Forwards selection – Proc Mean/GLMSELECT

Backwards elimination - GLMSELECT

  
Stepwise – GLMSELECT  


#### Model Metrics

The model developed was tested against other kaggle submissions and currently ranks in the 43rd percentile. The standard deviation of the final model is about 90,000$, with an r^2 of 0.92.

Of the forwards, backwards, and stepwise models used for predictive analysis, the \_\_\_\_\_\_\_\_\_ performed the best, with an r^2 of \_\_\_\_\_\_\_\_ and standard deviation of \_\_\_\_\_\_\_\_

#### Conclusion

The analysis supports the assumption that, all other factors being equal, sale prices in the Names neighborhood tend to be the highest of the three, followed by Edwards, then Brkside.   
The variables most strongly associated with changes in the final house sale price were (list variables with the most correlation?/influence? Here)

### 