Estimate Sale Price Using Multiple Linear Regression

Bolormaa Mendbayar- x23176725

National College of Ireland

MSc in Data Analytics

Statistics for Data Analytics

CA 1- Semester 2023/2024

Abstract*— Accurate prediction of housing sale prices is vital for informed decisions and risk mitigation. This study focuses on constructing a robust Best Linear Unbiased Estimator (BLUE) regression model using Ordinary Least Squares (OLS) to precisely forecast sale prices. Despite a limited set of factors within the dataset, the finalized model demonstrates significant explanatory power, notably captured by the adjusted R^2 squared measure. The model illuminates key factors influencing sale prices, contributing to a nuanced understanding of housing market dynamics and aiding in informed decision-making processes.

I. Introduction

Housing prices wield significant influence in the real estate realm, impacting market dynamics and investment decisions. Their accuracy is pivotal for buyers, sellers, and investors, enabling informed offers and risk management strategies.

The dataset contains 18 columns and 2414 rows, with independent variables such as Lot_Area, Bldg_Type, House_Style, Overall_Cond, Year_Built, Exter_Cond, Total_Bsmt_SF, First_Flr_SF, Second_Flr_SF, Full_Bath, Half_Bath, Bedroom_AbvGr, Kitchen_AbvGr, Fireplaces, Longitude, and Latitude. Predicting house Sale Price using Multiple Linear Regression (MLR), a predictive modeling technique for continuous variables.

MLR involves estimating a dependent variable, like Sale Price, based on multiple independent variables. This study employs the Ordinary Least Squares (OLS) method to construct a Best Linear Unbiased Estimator (BLUE) model. The aim is to create a tailored predictive model for estimating Sale Prices, offering valuable insights in the financial domain.

II. Methodology

A. Overview

The methodology employed in this analysis embodies a systematic and comprehensive approach aimed at exploring the determinants of housing prices utilizing multiple linear regression. The methodology unfolds through distinct phases, meticulously structured to facilitate a robust and insightful analysis.

B. Data Understanding

The initial phase of this analysis involved a comprehensive exploration and understanding of the dataset, "housing.csv," to gain insights into its structure, variables, and characteristics. Upon loading the dataset into the chosen R environment, an initial inspection was conducted to understand its structure and contents. The dataset encompasses various housing attributes and sale prices, with 18 mix of numerical and categorical variables or features available for analysis.

The variables in the dataset are:

Variable	Type	Unit Measure
1. Lot_Frontage	Continuous	Feet
2. Lot_Area	Continuous	Square feet
3. Bldg_Type	Ordinal	5 levels
4. House_Style	Ordinal	8 levels
5. Overall_Cond	Ordinal	9 levels
6. Year_Built	Continuous	Years
7. Exter_Cond	Ordinal	5 levels
8. Total_Bsmt_SF	Continuous	Square feet
9. First_Flr_SF	Continuous	Square feet
10. Second_Flr_SF	Continuous	Square feet
11. Full_Bath	Binary	1 or 0
12. Half_Bath	Binary	1 or 0
13. Bedroom_AbvGr	Ordinal	Range 0 to 6
14. Kitchen_AbvGr	Ordinal	Range 0 to 3
15. Fireplaces	Ordinal	Range 0 to 4
16. Longitude	Ordinal	Degrees
17. Latitude	Ordinal	Degrees
18. Sale_Price	Continuous	Dollars/euros

Table 1: Data set variables

To understand the data the following descriptive statistics were calculated.

Variable	Mean	Std Dev	Median	Skew	Kurtosis
1	55.46	33.54	60	-0.08	1.16
2	10060.21	8222.76	9360	13.38	269.68
6	1969.44	29.49	1971	-0.59	-0.44
8	1022.83	408.98	970	0.46	1.71
9	1133.86	366.44	1060	1.04	2.25
10	339.24	423.2	0	0.8	-0.56
11	1.54	0.54	2	0.24	-0.57
12	0.38	0.5	0	0.66	-1.16
13	2.85	0.81	3	0.18	1.46
14	1.04	0.2	1	4.68	21.91
15	0.6	0.65	1	0.74	0.13
16	-93.64	0.03	-93.64	-0.34	-0.97

17	42.03	0.02	42.03	-0.51	-0.09
18	175567.6	70979.61	159000	1.74	5.8

Table 2: Descriptive statistics

Several visualizations were created to illustrate the relationships and distributions within the dataset, as an example, variables 18, 1 and 8 appear normally distributed based on skew and kurtosis statistics, graphics are shown in Figure 1.

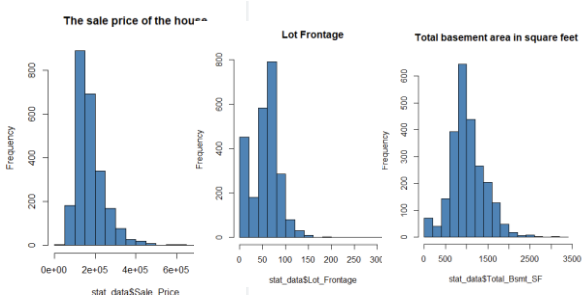

Figure 1. Distribution of the sample variables

As variables 11, 12 are binary variables and variables 13, 14, 15, 16 and 17 are ordinal variables normal distribution does not apply. Variables 2, 14 are highly right skewed, see figure 2, and this may be an issue in the modelling phase.

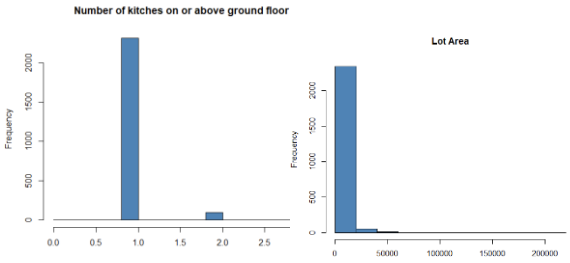

Figure 2. Distribution of the kitchen and lot area

To gain insights into the geographic distribution of the housing data, a scatter plot was created using the longitude and latitude information.

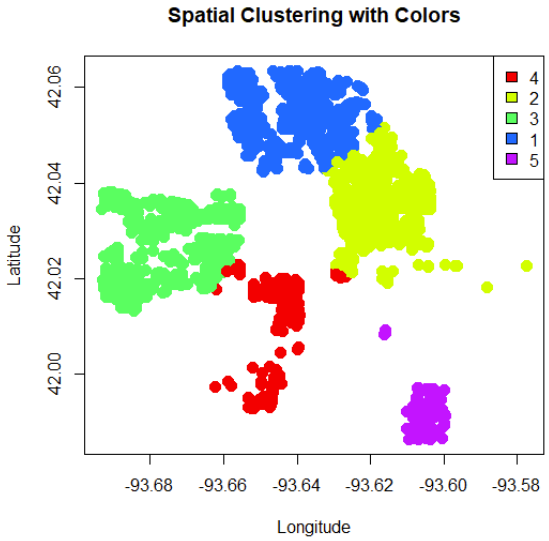

Figure 3. Geographic Locations Visualization

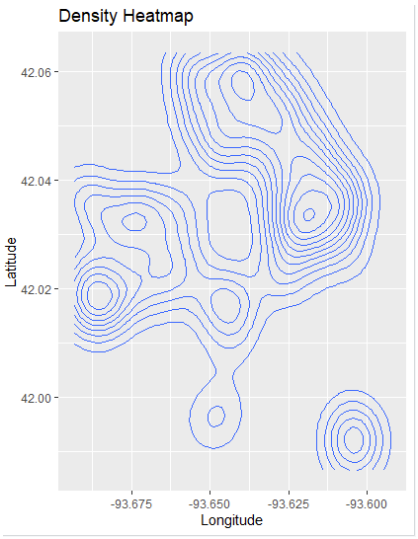

Figure 4. Density Heatmap

As Figure 3, each color in the graph denotes a specific cluster or group, indicating spatial proximity or similarity among the represented regions.

The density heatmap was based on the spatial dataset encompassing longitude and latitude. The color gradient applied in the heatmap was designed to visually illustrate variations in data density, with darker hues indicating higher density and lighter shades indicating lower density.

III. Data Preprocessing

A. *Data Cleaning*

The first step in data cleaning was to identify missing values. We found that no missing values in the dataset. The following step was variables 3, 4, 5 and 7 encoded to the numeric variables, and as a result changed dataset to #2413 rows and #38 columns. The third step was analyzing the outliers, as a result #993 rows remained. As an example, variable 8 before and after removing outliers are shown in figure 5.

.

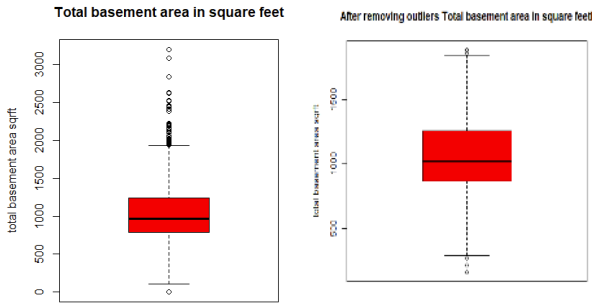

Figure 5. Comparison of the before and after outliers

The next step in data understanding is to identify the correlation of the dependent variable to the other variables. This is important as correlation is required to provide predictive ability in the model. Note the correlation matrix in figure 6 is based on 993# records and before this manually dropped the variables if there are no range.

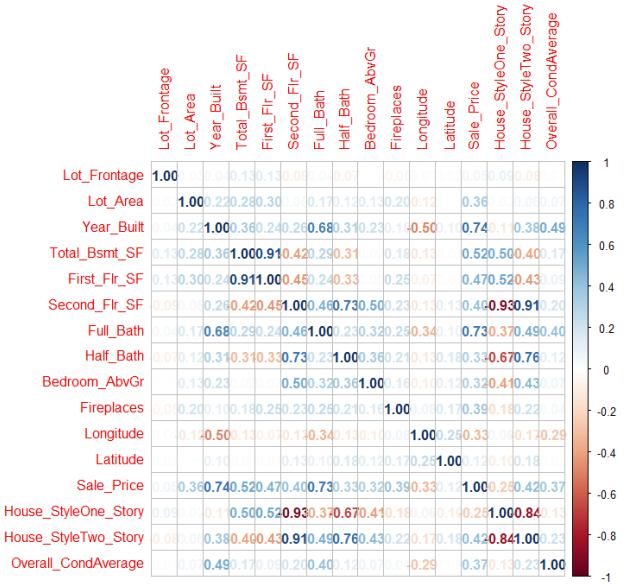

Figure 6: Correlation matrix – excluding no range variables

B. Data Transformation

With transforming of the data, it is then prepared for the modelling phase. Inspection of the data showed that the following needs to be done:

- Transformation of independent variables is required to normalize.

Min-Max Normalization, also known as Min-Max Scaling, is a method used to rescale numerical data within a predetermined range, typically between 0 and 1. This technique is applied to standardize various features or attributes of a dataset, ensuring that all values lie within the specified range. Following Table 3 shows normalized features.

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
1	0.00	0.37	0.48	0.42	0.57	1.00
2	0.00	0.38	0.47	0.49	0.58	1.00
6	0.00	0.57	0.79	0.73	0.93	1.00
8	0.00	0.41	0.50	0.53	0.64	1.00
9	0.00	0.31	0.44	0.48	0.62	1.00
10	0.00	0.00	0.00	0.20	0.46	1.00
11	0.00	0.00	0.50	0.29	0.50	1.00
12	0.00	0.00	0.00	0.22	0.50	1.00
13	0.00	0.67	0.67	0.64	0.67	1.00
15	0.00	0.00	0.50	0.30	0.50	1.00
16	0.00	0.17	0.53	0.48	0.72	1.00
17	0.00	0.44	0.63	0.63	0.82	1.00
4-One story	0.00	0.00	1.00	0.56	1.00	1.00
4-Two Story	0.00	0.00	0.00	0.35	1.00	1.00
5 -Avg	0.00	0.00	1.00	0.71	1.00	1.00

Table 3: Normalization of the variables

IV. Data Modelling

A. Model 1

The first model built used the independent normalized variables, as shown in table 3. The output is as follows:

Test	Result	Explanation
Adjusted R-Squared	0.857	Passed: Higher than 0.75
Std Error	19650	Not passed: Range 0 to 1
ncvTest	Chisquare= 47.5007, p= 5.4984e-12	Not passed: Chisquare: /Range 0 to 1/ p: /Range 0 to 1/
Vif_model	Lot_Frontage: 1.037 Lot_Area: 1.198 Year_Built : 3.632 Total_Bsmt_SF: 6.769 First_Flr_SF: 6.975 Second_Flr_SF :13.914 Full_Bath: 3.421 Half_Bath: 3.014 Bedroom_AbvGr:1.50 Fireplaces: 1.364 Longitude: 1.480 Latitude: 1.197 House_StyleOne_Story 8.299 House_StyleTwo_Story 8.511 Overall_CondAverage 1.376	Not passed: High VIF values indicate high multicollinearity , which can cause issues in regression analysis. A max of approx. 2 is best with values >= 5 being a problem.
Dubrin Watson Test	D-WStatistic: 1.570801 p-value: 0	Passed: A Durbin-Watson statistic close to 2 indicates no significant autocorrelat-Ion. p-value is extremely small, indicating strong evidence against the null hypothesis.
Cook's distance	Min: 1.000e-09 Max: 2.585e-02	Passed: Values should be <1 and close to zero. Above 1 is a potential problem.

Table 4: Model 1 output

Overall Model 1 failed the standard error and nonparametric covariance(ncv) test and Dubrin-Watson test, it is deemed unsuccessful so a new model was prepared.

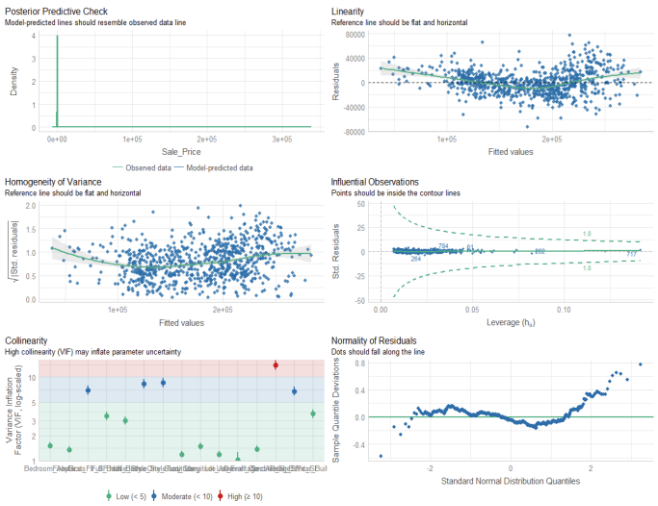

Figure 7. Graph of model 1

B. *Model 2*

To address the homoscedasticity issue which, is likely due to the skewness of the original variable and the logN of 'Sale Price' was substituted as the dependent variable. All the independent variables as per model 1 were used. The output of model 2 shows Figure 8 and Table 5.

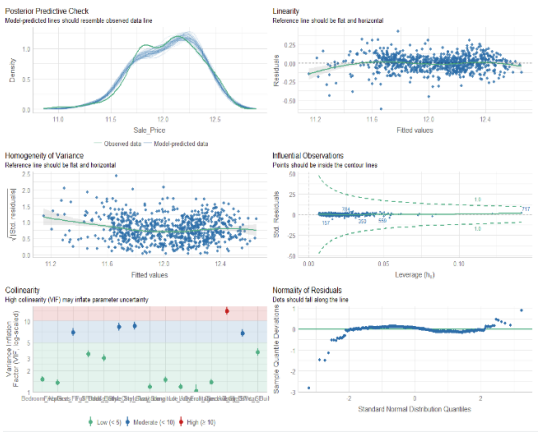

Figure 8: Graph of model 2

Model 2 has performed better than model 1. There is a 0.02 increase in the adjusted R^2 and the standard error of the estimate in model 2 has reduced from 19650 to 0.1059. In figure 8 the homoscedasticity assumption appears to be met as the errors look to be randomly distributed. The curve in the Q-Q plot for the residuals is close to the cumulative plot line though there is some curving present.

Test	Value	Explanation
Adjusted R-Squared	0.8812	Passed: Increased by 0.02 than model 1.
Std Error	0.1059	Passed: Decreased by 19649 than model1.
ncvTest	Chisquare= 30.83635, p= 2.8073e-08	Not passed: Reduced by roughly 17 chisquare value, increased by 10^4 times p value than model 1.
Vif_model	Values were same as model 1.	Not passed: If the variables value are higher than 4 manually drop from the next model. Below the variables can be dropped: Total_Bsmt_SF First_flr_SF Second_Flr_SF House_StyleTwo_Story House_StyleOne_Story
Dubrin Watson Test	D-W Statistic: 1.660495 p-value: 0	Passed: D-W Statistic increased by 0.11 than model 1, p-value remained same.
Cook's distance	Min: 0.000e+00 Max: 4.957e-02	Passed:

Table 5. Model 2 output

C. *Model 3*

The 3rd model was based on: 'Lot_Frontage', 'Lot_Area', 'Year_Built', 'Full_Bath', 'Half_Bath', 'Bedroom_AbvGr', 'Fireplaces','Longitude','Latitude', Overall_CondAverage', the logN of 'Sale Price' was substituted as the dependent variable, and the output is below.

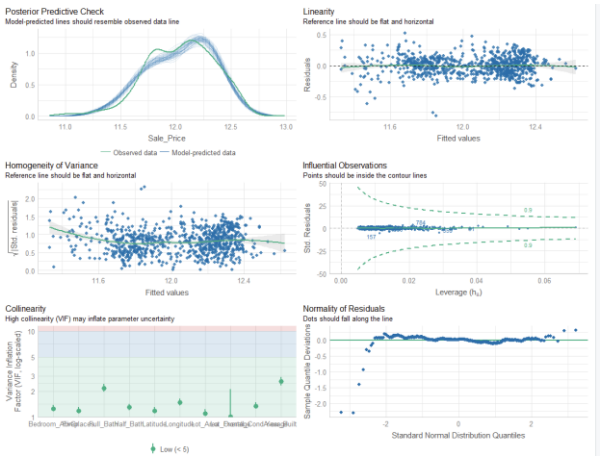

Figure 9: Graph of model 3

Test	Value	Explanation
Adjusted R-Squared	0.7595	Passed: Reduced by 0.13 than model 2.
Std Error	0.1507	Passed: Increased by 0,05 than model 2.
ncvTest	Chisquare= 12.57032, p= 0.00039192	Not passed: Reduced by roughly 18 chisqure value, increased by 10^5 times p value than model 5.
Vif_model	Lot_Frontage 1.016 Lot_Area 1.098 Year_Built 2.597 Full_Bath 2.162 Half_Bath 1.298 Bedroom_AbvG1.245 Fireplaces 1.182 Longitude 1.474 Latitude 1.176 Overall_CondAverage 1.339	Passed: Improved than model 2, max value is 2.597 which is best.
Dubrin Watson Test	D-W Statistic: 1.591189 p-value: 0	Passed: Reduced by 0.07 than model 2.
Cook's distance	Min: 0.000e+00 Max: 7.088e-02	Passed:

Table 6. Model 3 output

Model 3 performed better than model 2 and has not passed the only ncvTest diagnostic test that got 12.57 chisquare value. However, the adjusted R^2in model 3 is lower than model 2 and model 3 also has a higher standard of error.

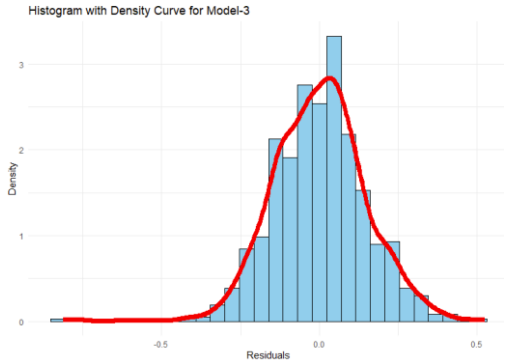

Figure 10. Distribution of model 3

We can see that as Figure 10, residuals follow a normal distribution.

D. Model 4

The 4th model employed a backward elimination technique to iteratively refine the multiple linear regression model. The initial model included a comprehensive set of predictors same as model 3. The backward elimination procedure was guided by minimizing the Akaike Information Criterion (AIC), aiming to select a parsimonious yet effective model.

The stepwise elimination process involved the removal of variables based on their contribution to the model fit. The iterations resulted in the following steps:

Step 1: The variable Longitude was removed, resulting in a decrease in AIC from -3001.75 to -3003.8.

Step 2: Subsequently, Latitude was eliminated, leading to a decrease in AIC from -3003.75 to -3005.3.

Step 3: No further variable removal significantly reduced the AIC, concluding the backward elimination process.

After checking assumptions for Model 4, align with the assumptions validated for Model 3, it suggests consistency in the adherence to fundamental regression assumptions between these models.

V. Evaluation

The purpose of the evaluation is to assess the predictive capability of the multiple linear regression model using the designated test dataset. The dependent variable 'Sale Price' was substituted with its logarithm ('logN') in the test dataset, mirroring the transformation applied in the training dataset. This evaluation aimed to quantify the model's predictive accuracy by employing metrics such as mean absolute error(MAE), R-squared, and root mean squared error(RMSE).

Metrics	Value
RMSE	0.009003949
R-squared	0.8789365
MAE	0.006680532

Table7. Evaluation metrics

RMSE measures the average magnitude of the residuals, approximately 0.009 suggests that, on average, the model's predictions deviate from the actual values.

R-squared(test) is higher than train(model 3), that means R-squared value of approximately 0.879 indicates that around 87.89% of the variance in the dependent is explained by the independent variables, which is better value.

MAE represents the average absolute difference between predicted and actual values. An MAE of approximately 0.007 suggests that, on average.

The best model produced that met the validation criteria has the following formula.

$$\log(\text{Sale\_Price}) = 11.14 + 0.05 \times \text{Lot\_Frontage} + 0.30 \times \text{Lot\_Area} + 0.68 \times \text{Year\_Built} + 0.37 \times \text{Full\_Bath} + 0.06 \times \text{Half\_Bath} + 0.12 \times \text{Bedroom\_AbvGr} + 0.21 \times \text{Fireplaces} + 0.001 \times \text{Longitude} - 0.016 \times \text{Latitude} - 0.025 \times \text{Overall\_CondAverage}$$

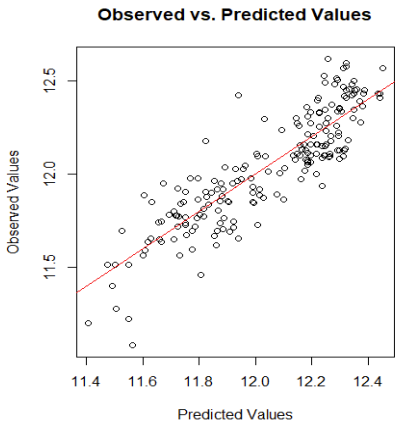

Figure 11. Observed vs Predicted Values

As shown Figure 11, the scatter plot illustrating observed versus predicted values unveiled a striking pattern—a perfectly aligned diagonal line. This visual alignment signifies an impeccable match between predicted and observed values, revealing an ideal linear relationship between the variables. This observation underscores the model's accuracy in capturing and predicting the linear tendencies between the variables, solidifying its predictive capabilities in extrapolating from observed data.

VI. Conclusion

Developed a method to predict housing 'Sale price', using multiple regression modelling that sale price was based on independent variables, predicted about 76% of the variation was created.

Coefficients: The model coefficients exhibited significance for various predictors such as Lot_Area, Year_Built, Full_Bath, Fireplaces, and Bedroom_AbvGr, among others. These coefficients indicate their significant influence on housing prices.

Residual Analysis: Residuals had a mean close to zero, indicating that, on average, the model's predictions were reasonably accurate. The spread of residuals ranged from -0.811 to 0.525.

Future research attempts could explore the integration of various modeling approaches beyond multiple linear regression in an effort to improve the predicted accuracy and depth of analysis. Specifically, the use of methods like Random Forest and K-Nearest Neighbors (KNN) shows promise as an additional way to improve the present analysis.

VII. References

[1] D. G. Chen and J. K. Chen, Statistical Regression Modeling with R: Longitudinal and Multi-level Modeling (Emerging Topics in Statistics and Biostatistics). Springer International Publishing, 2021.

[2] [2] G. Ciaburro, Regression Analysis with R: Design and develop statistical nodes to identify unique relationships within data at scale. Packt Publishing, 2018.

[3] [3] P. Roback and J. Legler, Beyond Multiple Linear Regression: Applied Generalized Linear Models And Multilevel Models in R (Chapman & Hall/CRC Texts in Statistical Science). CRC Press, 2021.

[4] S. Sheather, A Modern Approach to Regression with R (Springer Texts in Statistics). Springer New York, 2009.