PROPERTY CHARACTERISTICS AND SELLING PRICES IN U.S. REAL ESTATE

FIN-403 Econometrics

Assignment 1
Group 12

Professor
Prof. Andreas Fuster

Teaching Assistant
Federico Baldi Lanfranchi

Authors
Charalampos Salis (367920)
Ribeiro de Carvalho Paulo (314985)

October 23, 2023

(a)

Duplicates of specific property identifiers (`pid`) have been identified and after their removal, the data size is reduced from 3004 to 2930 unique transactions (*Appendix :* Table 3), revealing that 74 transactions were duplicates.

(b)

The categorical attribute informing about the presence of a central air conditioning use *textual* format, so we use one-hot encoding. For numerical data we use measures of central tendency, dispersion and shape of distribution. We also consider date as numerical, since discrete variables with a large number of categories converge to a continuous variable.

Table 1: Continuous Descriptive Statistics

Variable	mean	median	mode	min	max	sd	cv	iqr	skew	kurt
saleprice	180693.67	160000.0	135000	12789	755000	79956.59	44.25	84212.5	1.74	5.10
grlivarea	1500.71	1442.0	864	334	6642	513.82	34.24	616.75	1.60	8.02
yearbuilt	1971.36	1973.0	2005	1872	2010	30.25	1.53	47.0	-0.60	-0.50

Both `saleprice` and `grlivarea` variables appear to follow right-skewed distributions, which implies the presence of extreme values. To reinforce this observation, we have generated histograms (*Appendix :* Figure 5, 6) for both attributes. Additionally, the distribution of the `yearbuilt` is skewed, since most houses were built in the recent years (*Appendix*, Figure 7). Such skewed distributions lead to biased estimates and issues in linear regression analysis. The presence of extreme values can distort the interpretation of summary statistics and the predictive performance of regression models.

For the rest of the categorical variables, the frequency tables (python file) imply imbalance between the different classes (can be problematic for classification problems).

(c)

We constructed two models (2^{nd} model in R file) to visually demonstrate the impact of the three outliers (Figure 1). The final model with the outliers is:

$$\mathbb{E}[saleprice_i] = 21180.3 + 106.3 grlivarea_i$$

Figure 1: Models fitted

Coefficient β_2, for both models, shows that the relation between `saleprice` and `grlivarea` is statistically significant and positive (*Appendix*, Tables 4, 5), implying that for the increase in one unit of the living space,

we expect the price of the property to increase by 106.3 monetary units. The three outliers have abnormal `salecondition`. These observations have significantly large leverage scores (*Appendix*, Table 6).

(d)

The logarithmic transformation is expected to reduce the skewness of the distribution an bring it closer to normality (which is also implied by Figure 2). From this preferred distribution of `saleprice`, the new model is:

$$\mathbb{E}[log(saleprice_i)] = 11.1765 + 0.0006 grlivarea_i$$

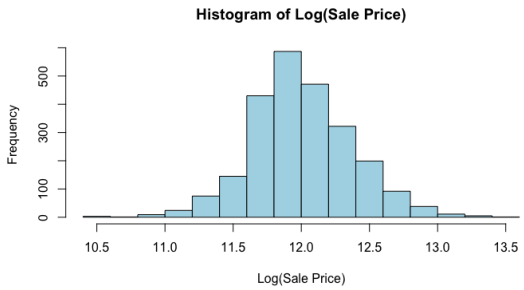

Figure 2: Histogram of log(`saleprice`)

Finally, the predicted value of a property with 2,500 square feet is $\$299,803.42 (\approx e^{prediction+0.5\times var(\epsilon)} = e^{12.58+0.5\times 0.0641})$.

(e)

To check for homoskedasticity, we examine a plot of residuals. Residuals should be uniformly distributed around zero. However, the non-logarithmic OLS model exhibits signs of heteroskedasticity (Figure 3). On the other hand, the logarithmic model does not present significant signs of heteroskedasticity (Figure 4).

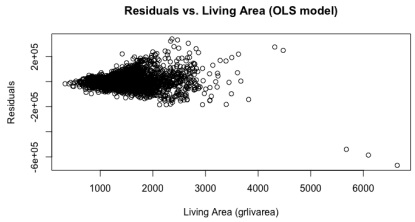

Figure 3: Residuals of OLS model

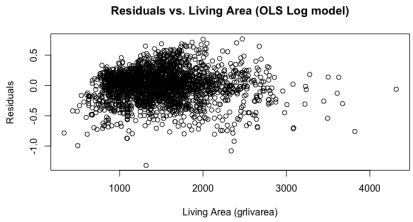

Figure 4: Residuals of OLS Logarithmic model

(f)

The more complex model has the following form (see also *Appendix*, Table 7):

$$\mathbb{E}[\log(saleprice_i)] = 0.7665 + 0.0005 grlivarea_i + 0.0053 yearbuilt_i - 0.0285 fullbath_i + 0.2043 centralair_i$$

The intercept (β_1) represents the expected value of the natural logarithm of sale price (ln(saleprice)) when all the other predictor variables (grlivarea, yearbuilt, fullbath, and centralair) are equal to zero. Since this is unrealistic, we cannot practically interpret it more than simply being the intercept in the regression formula

(since a house with zero features is no house).

The coefficient β_5 implies that the presence (or absence respectively) of central air-conditioning positively affects the logarithmic prices. The numerical value of β_1 is only useful for model fitting, i.e., since *centralair* is categorical and not numerical, the increase of one unit cannot imply expected changes at the price of the property.

For the final part of the question, we test the hypothesis:

$$H_0 : 100\beta_2 - 9\beta_3 = 0$$

With p-value: 0.076 we conclude that there is positive evidence that the effects of the two changes are equivalent.

(g)

The collinearity of the exogenous variables can be checked using the Variance Inflation Factor (`VIF`).

Table 2: `VIF` Summary of Attributes

	grlivarea	yearbuilt	fullbath	centralair
`VIF`	1.77	1.45	2.10	1.17

None of the attributes have significant `VIF` (>5) scores, implying that there is no evidence of collinearity.

For the `RESET` test, we obtained a `p-value` of `0.5242`, indicating that we lack sufficient confidence to assert a significant improvement in the model. However, if the abnormal observations were not removed, the `p-value` would have been 2.2×10^{-16}, concluding that adding higher-order attributes would enhance the model.

(h)

By adding the variable *overallqual*, the β_2 coefficient is reduced from 0.0005 to 0.0004, thus the impact of *grlivarea* on the *log*(*saleprice*) is reduced.

A possible explanation is that *overallqual* and log(*saleprice*) have a correlation value of approximately 0.81, indicating a strong positive linear correlation between them, while the correlation between *overallqual* and *grlivarea* is 0.55. Hence, *log*(*saleprice*) is now less influenced by *grlivarea*.

(i)

From the new model (*Appendix*, Table 8), we conclude that *oldhouse* is statistically insignificant, thus the old houses do not influence more/less the *log*(*saleprice*) compared to the newer ones and the interaction term is statistically significant, which implies that the impact of living area on the sale price is different for old houses compared to newer ones.

(j)

We perform the Chow test by fitting two regression models for each partition of the dataset and then a general model to the entire dataset. Through the use of the SSR values of each model, we can construct the Chow statistic and perform the test. In this case, the p-value of the test is ≈ 1, which implies that there is no structural break in the data (which is also implied by the fact that the *oldhouse* coefficient was statistically insignificant in (i)).

Appendix

(a)

Table 3: Unique Transactions Characteristics

id	pid	overallqual	yearbuilt	centralair	grlivarea	fullbath	salecondition	saleprice
1	526301100	6	1960	Y	1656	1	Normal	215,000
...	...	...	...	...	...	...	...	...
2930	924151050	7	1993	Y	2000	2	Normal	188,000

(b)

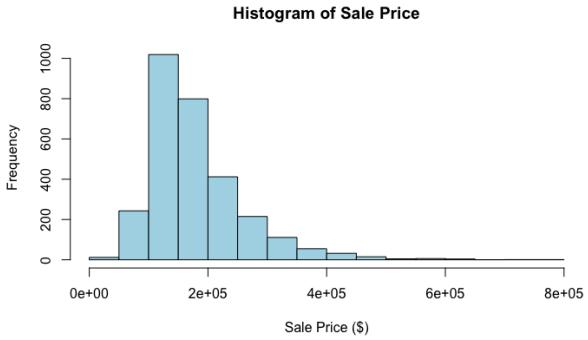

Figure 5: Histogram of `saleprice`

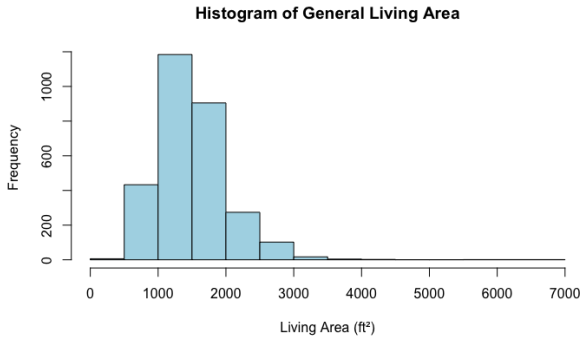

Figure 6: Histogram of `grlivarea`

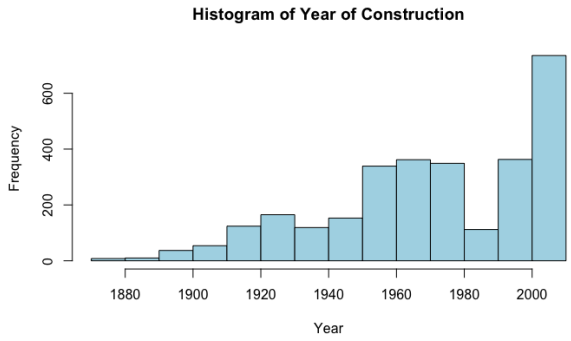

Figure 7: Histogram of `yearbuilt`

(c)

Table 4: OLS Estimators w/ outliers

	Estimate	Std. Error	p-value
(Intercept)	21180.3	3331.6	2.37×10^{-10}
grlivarea	106.3	2.1	$< 2 \times 10^{-16}$

Table 5: OLS Estimators w/o outliers

	Estimate	Std. Error	p-value
(Intercept)	4091.200	3247.361	0.208
grlivarea	118.124	2.062	$< 2 \times 10^{-16}$

Table 6: Leverage Score of Suspicious Observations

`pid`	908154235	908154195	908154205	`mean score`
`leverage score`	14.92001×10^{-4}	13.69128×10^{-4}	12.750×10^{-4}	3.413×10^{-4}
`high leverage`	TRUE	TRUE	TRUE	-

(f)

Table 7: OLS Estimators with log(`saleprice`)

	Estimate	Std. Error	t-error	p-value
(Intercept)	7.665×10^{-1}	2.965×10^{-1}	2.586	0.00978
grlivarea	5.019×10^{-4}	1.047×10^{-5}	47.928	$< 2 \times 10^{-16}$
yearbuilt	5.255×10^{-3}	1.554×10^{-4}	33.813	$< 2 \times 10^{-16}$
fullbath	-2.849×10^{-2}	1.013×10^{-2}	-2.813	0.00495
centralair	2.043×10^{-1}	1.689×10^{-2}	12.094	$< 2 \times 10^{-16}$

(i)

Table 8: OLS Estimators of model with interaction

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.100×10^{1}	2.838×10^{-2}	387.462	$< 2 \times 10^{-16}$
grlivarea	5.363×10^{-4}	1.374×10^{-5}	39.046	$< 2 \times 10^{-16}$
fullbath	1.211×10^{-2}	1.111×10^{-2}	1.090	0.276
centralair	3.169×10^{-1}	1.782×10^{-2}	17.785	$< 2 \times 10^{-16}$
oldhouse	-2.631×10^{-2}	2.815×10^{-2}	-0.935	0.350
grlivarea:oldhouse	-1.323×10^{-4}	1.794×10^{-5}	-7.376	2.23×10^{-13}