Hedonic Analysis of House Prices

The data set "house.dta" contains information on recent sale prices and characteristics of houses in four California cities: Davis, Fresno, Berkeley, and San Francisco. You can see all the variable definitions by using the "describe" command in Stata. You will use these data to construct a hedonic price model for house sales in these cities. You can find more information about hedonic analysis in the document "Hedonic Price Models."

The 'Zestimate' Model

The Zestimate model does a good job in predicting the actual sale prices of houses. The R-squared is 0.9753, which signifies that the regression model shows that the Zestimates are able to explain 97.53% of the variability in house prices. the standard error was only .0113945, which is very close to 0. Thus, this indicates that actual value and predicted values are close to each other therefore Zillow is able to closely estimate the true value of housing prices

```
reg Price Zestimate
```

Source	SS	df	MS	Number of obs	=	225
				F(1, 223)	=	8807.08
Model	1.7584e+15	1	1.7584e+15	Prob > F	=	0.0000
Residual	4.4523e+13	223	1.9965e+11	R-squared	=	0.9753
				Adj R-squared	=	0.9752
Total	1.8029e+15	224	8.0486e+12	Root MSE	=	4.5e+05

Price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Zestimate	1.069332	.0113945	93.85	0.000	1.046878	1.091787
_cons	-107317	34240.42	-3.13	0.002	-174793.2	-39840.84

Price = b0 + b1*(Zestimate)

Is the Zestimate a good predictor of housing prices

Testing, at the 99% significance level, if the zestimate a good predictor of realized house prices.

H0: b1 = 1
Ha b1 ≠ 1
Prob > F = 0.0000
0.0000> (alpha = .01)

Reject the null hypothesis therefore with a 99% significance level zestimates are not able to exactly predict the realized hose price. Therefore, we conclude that Zestimates by Zillow does not do a good job predicating realized house prices.

Hedonic Housing Model:

Ln(price) = b0 + b1*ln(size) + b2*ln(rooms) + b3*ln(lotsize) + b4*ln(year) +b5 *ln(baths)+ b6* (parking) + b7* (type) + b8* Berkeley + b9*Davis
+ b10* Fresno

In order for my ordinary least square model to stay BLUE (Best Linear Unbiased Estimate) we have to assume CR1, CR2, and CR3: For CR 1, we have to make the assumption that our sample is representative of the population of the population that we are testing, and if that fails then it would cause my model to lose the U in BLUE. For CR 2, we need to have homoscedastic error, meaning that the variance is constant amongst all the data. On top of that we also have to assume CR. CR 3 refers to how the errors have to be uncorrelated across all the observations. If CR 2 or CR 3 is broken then my model will lose the B in BLUE. CR 4 does not have to be assume because the data has large sample size.

Hedonic Pricing Regression Model

```
reg lnprice lnsize lnrooms lnbaths lnlotsize lnyear Parking Type i.Berkeley i.Davis i.Fresno
```

Source	SS	df	MS	Number of obs	=	225
				F(10, 214)	=	180.42
Model	179.507539	10	17.9507539	Prob > F	=	0.0000
Residual	21.2920702	214	.099495655	R-squared	=	0.8940
				Adj R-squared	=	0.8890
Total	200.799609	224	.896426828	Root MSE	=	.31543

lnprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnsize	.8034791	.0832728	9.65	0.000	.6393391	.967619
lnrooms	-.3373936	.0919414	-3.67	0.000	-.5186204	-.1561669
lnbaths	.2986587	.0791981	3.77	0.000	.1425505	.454767
lnlotsize	.1228785	.047293	2.60	0.010	.0296587	.2160983
lnyear	.3347453	1.706653	0.20	0.845	-3.029258	3.698749
Parking	.066535	.0726392	0.92	0.361	-.0766449	.209715
Type	-.0380212	.0771806	-0.49	0.623	-.1901527	.1141103
1.Berkeley	-.4914354	.0648378	-7.58	0.000	-.6192379	-.3636329
1.Davis	-1.141518	.0833804	-13.69	0.000	-1.30587	-.9771658
1.Fresno	-1.954008	.0853338	-22.90	0.000	-2.122211	-1.785806
_cons	5.054129	12.92305	0.39	0.696	-20.41863	30.52689

Interpretation of the Coefficients:

b0, the intercept, in this model is equal to 5.054129 This means that lnprices will equal to 5.054129 if all the other variables were equal to 0.

b1 here equals .8034791 which means for the variable size, a 1% increase in the size of a house leads to a .8034791% ***increase*** in price

b2 here equals to -.3373936. This means here for rooms a 1% *increase* in rooms causes a .3373936% ***decrease*** in price. This relationship between price and rooms doesn't make sense since usually the more rooms a house has the higher the sale price.

b3 equals .1228785. This means a 1% *increase* in lot size causes .1228785 % ***increase*** in price. The relationship between lot size and price does make sense here because usually a bigger lot means that the house should be worth more.

b4 equals to .3347453 This means 1% increase in year causes .3347453 % ***increase*** in price. The relationship between price and years makes sense since an older house is often priced less compared to a newer house.

b5 equals to .066535. Therefore a 1% *increase* in baths causes a .066535 % ***increase*** in price. The relationship between price and number of baths makes sense, since more bathrooms usually means a house is worth more.

b6 equal to .066535. This means for parking, having an at home parking cause price to ***increase*** by .066535%. This relationship makes sense since having a parking at home is highly valued and therefore can increase the price of a house with parking.

b7 equals -.0380212. This means for type of home, having condo over a house cause a .0380212 % ***decrease***. This relationship between type of house and price makes sense since condos are usually smaller they are also usually cheaper.

b8 equals -.4914354 and if the house located in Berkeley instead of SF the price will ***decrease*** by .4914354%.

b9 equals -1.141518 and if the house located in Davis instead of SF the price will ***decrease*** by 1.141518 %.

b10 equals -1.954008 and if the house located in Fresno instead of SF the price will ***decrease*** by 1.954008 %.

R-square Interpretation:

R square = 0.8940 therefore my hedonic price regression model explains 89.40% of the variation in the data. I believe this is a good model since my model helps explain the majority of the variation in the data.

Hypothesis Test:
Testing the null hypothesis that, controlling for other characteristics of houses, house prices are the same in all four cities, at the 99% significance level.

Ho: b8 = 0 and b9 = 0 and b10 = 0
Ha b8 ≠0 or b9 ≠ 0 or b10 ≠ 0
Wald Stat = (0.8940- 0.6125) / ((1- 0.8940)/(225-10-1)) = 568.311320755

Using a Chi-square table the chi-squared stat at alpha = 0.05 and 3 degrees of freedom 1 is 9.21

568.311320755 > 9.21
Therefore, we reject the null hypothesis, Ho, and with a 99% significance level there is evidence to suggest that the housing prices in the 4 cities are not the same. This makes sense since certain cities are more desirable to live in, therefore have a higher housing price.

Confidence Intervals:

95% confidence interval around the value of an additional square foot of house size.
Confidence interval = b1 +/- standard error of b1 * $t_{.05, 225-10-1}$
[.8034791 - .0832728 *1.96 , .8034791 + .0832728 *1.96]
The confidence interval for each additional square footage is **[.6393391, .967619]**

We are 95% confident that the true b1, value of an additional square foot in a house, fall within the confidence interval **[.6393391, .967619]**

Testing whether an additional square foot add the same to the sales price of a house in all four cities? Show how you would modify your model to test this, and report the results. Does allowing an additional square food to add different amounts to housing prices result in a better model? Please test this at the 95% confidence level.

Ln(price) = b0 + b1*ln(size) + b2*ln(rooms) + b3*ln(lotsize) + b4*ln(year)
+b5 *ln(baths)+ b6* (parking) + b7* (type) + b8* Berkeley + b9*Davis
+ b10* Fresno + b11*Berkeley*lnsize + b12*Davis*lnsize + b13*Fresno*lnsize

Ho: b11 = 0 and b12 = 0 and b13 = 0
Ha b11 ≠0 or b12 ≠ 0 or b13 ≠ 0
R^2_{alt} = 0. 9183
R^2_{null} = 0. 8940
Wald Stat = (0.9183- 0.8940) / ((1- 0.9183)/(225-13-1)) = 62.7576499

Using a Chi-square table chi-squared stat at alpha = 0.05 and 13 degrees of freedom is 7.82

62.7576499> 7.82

We reject the Ho and at the 95% significance level we conclude there is evidence supporting that an increase in square footage does not add the same value for each of the cities. This makes sense due to the fact that people value an increase in square footage in certain more desirable cities, such as San Francisco, compared to other cities.

Allowing additional square foots to adding different amounts to housing prices does result in a better model as denoted by a higher R-squared. Now the model is able to explain 91.83% of the variability in the data.

Is the model created a good estimator?

Source	SS	df	MS	Number of obs	=	225
				F(10, 214)	=	180.42
Model	179.507539	10	17.9507539	Prob > F	=	0.0000
Residual	21.2920702	214	.099495655	R-squared	=	0.8940
				Adj R-squared	=	0.8890
Total	200.799609	224	.896426828	Root MSE	=	.31543

Price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Yestimate	1	.025644	39.00	0.000	.9494644	1.050536
_cons	-.0225298	77655.3	-0.00	1.000	-153032.1	153032.1

Hypothesis Test:

Ho Yestimate = 1
Ha Yestimate $\neq$ 1

```
. test Yestimate==1

 ( 1)  Yestimate = 1

       F(  1,   223) =    0.00
            Prob > F =    1.0000
```

Based off the hypothesis test, the housing model about is a good fit and does a good job predicting housing prices.

Conclusion:

In this analysis, it is learned that multiple variables affect housing prices such as house size, location, year built etc. It is noteworthy that location plays a large factor in house prices. House in locations such as San Francisco and Berekely were sold at a higher price compared to a house in Fresno and Davis. This indicates that some locations are more desirable then others and people are willing to pay more for a certain location. Size was another house major contributing factor to house prices. Houses that were larger had overall were more expensive