Linear Regression on House Prices

Presented by:
Nitish Sou

DATA SUMMARY AND PRE-PROCESSING

- Our data contains the following variables: age, no. of neighbourhoods, no of rooms, area in sqft, no of bathrooms and Y81.
- Necessary data pre-processing techniques have been applied to ensure proper data analysis, like replacing null value with mean value of the variable.
- **Challenge:** Our main aim is to analyse the house price on the basis of age, nbh (no. of neighbourhoods), rooms (no. of rooms), area (sqft), baths (no. of bathrooms) and Y81 (dummy variable).

	age	nbh	price	rooms	area	baths	y81
1.	48	4	60000	7	1660	1	0
2.	83	4	40000	6	2612	2	0
3.	58	4	34000	6	1144	1	0
4.	11	4	63900	5	1136	1	0
5.	48	4	44000	5	1868	1	0
6.	78	4	46000	6	1780	3	0
7.	22	4	56000	6	1700	2	0
8.	78	4	38500	6	1556	2	0
9.	42	4	60500	8	1642	2	0
10.	41	4	55000	5	1443	2	0

Variable	Obs	Mean	Std. Dev.	Min	Max
age	321	18.00935	32.56585	0	189
nbh	321	2.208723	2.164353	0	6
price	321	96100.66	43223.73	26000	300000
rooms	321	6.58567	.9012042	4	10
area	321	2106.729	694.9579	735	5136
baths	321	2.339564	.7705265	1	4
y81	321	.4423676	.4974428	0	1

LINEAR-LINEAR MODEL

HOUSE PRICE = B0 + B1 (AGE) + B2 (NBH) + B3 (CBDDIST) + B4 (INTSTDIST) + B5 (ROOMS) + B6 (AREA) + B7 (BATHS) + B8 (DISTINCINERATOR) + B9 (Y81) + ERROR

Based on Regression Result:

House Price = -20604.4 - 216.1286 (age) - 2259.32 (NBH) -.5689782 (cbddist) + 0.251932 (intstdist) + 3988.193 (rooms) + 22.04232 (area) + 13491.17 (baths) + 0.5106087 (distincinerator) + 35275.52 (y81) + Error

Because we are fitting a linear-linear model, we are assuming that the relationship really is linear, and that the errors, or residuals, are simply random fluctuations around the true line.

Source	SS	df	MS			
Model	4.2182e+11	9	4.6869e+10	Number of obs	=	321
Residual	1.7603e+11	311	566016448	F(9, 311)	=	82.81
				Prob > F	=	0.0000
Total	5.9785e+11	320	1.8683e+09	R-squared	=	0.7056
				Adj R-squared	=	0.6970
				Root MSE	=	23791

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-216.1286	49.94408	-4.33	0.000	-314.3996	-117.8576
nbh	-2259.32	659.1081	-3.43	0.001	-3556.195	-962.4448
cbddisttocentralbus	-.5689782	1.847156	-0.31	0.758	-4.203481	3.065524
intstdistance	.251932	1.297199	0.19	0.846	-2.300464	2.804328
rooms	3988.193	1970.35	2.02	0.044	111.2908	7865.095
area	22.04232	2.809599	7.85	0.000	16.51409	27.57054
baths	13491.17	2896.641	4.66	0.000	7791.679	19190.66
distfromhousetoincinerator	.5106087	.7697863	0.66	0.508	-1.004039	2.025257
y81	35275.52	2806.53	12.57	0.000	29753.34	40797.71
_cons	-20604.4	11984.16	-1.72	0.087	-44184.7	2975.89

REGRESSION AND ITS INTERPRETATION

- Keeping all other factors constant, with every 1 year increase in the age of a house, the price of the house increases by -216 dollars.
- Keeping all other factors constant, increase in number of rooms and area (sqft), the house price increases by $3988.193 and $22.042.
- Keeping everything else constant, the difference in average house price before and after 1981 is $35275.52.

CHECKING THE MODEL PERFORMANCE

R2 is equal to 0.07056, having a high R2 value shows that the proportion of the variation in the dependent variable that is explained by the independent variables is quite high.

The value of Prob(F) is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero). Since Prob(F)<0.05, we can reject the null hypothesis.

CHECKING FOR INSIGNIFICANT VARIABLE

From the regression table we can see that central bus distance, inter-state distance and distance from incinerator have P-values exceeding the significant level of 0.05, therefore they're insignificant.

```
. test age=nbh=rooms=area=baths=y81=0

 ( 1)  age - nbh = 0
 ( 2)  age - rooms = 0
 ( 3)  age - area = 0
 ( 4)  age - baths = 0
 ( 5)  age - y81 = 0
 ( 6)  age = 0

       F(  6,   314) =  124.66
            Prob > F =    0.0000
```

PERFORMING REGRESSION AFTER REMOVING THE INSIGNIFICANT VARIABLE

After removing the insignificant variable, the new performance parameters we get are:

R2_new = 0.7043

Prob(F)_New < 0.05

There is a decrease in new R2 from 0.7056 to 0.7043 after omitting insignificant variables. This is a common phenomenon as increase in variable features will always slightly increase the R2, but never decrease it.

CONCLUSION: AFTER REMOVING THE INSIGNIFICANT VARIABLES (, R2 VALUE DECREASED ONLY SLIGHTLY. THIS INDICATES THAT THE OVERALL USEFULNESS OF THE MODEL DECREASED ONLY SLIGHTLY. THERE WAS NOT MUCH IMPACT ON THE OVERALL PERFORMANCE OF MODEL, THAT WE GOT WITH ORIGINAL MODEL.

```
. reg price age nbh rooms area baths y81, vce(robust)

Linear regression                               Number of obs =      321
                                                F(  6,   314) =   116.74
                                                Prob > F      =   0.0000
                                                R-squared     =   0.7043
                                                Root MSE      =    23727
```

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-225.5894	54.6436	-4.13	0.000	-333.1033	-118.0755
nbh	-2109.249	618.7794	-3.41	0.001	-3326.727	-891.7707
rooms	4251.945	1725.201	2.46	0.014	857.5288	7646.361
area	21.65402	5.567451	3.89	0.000	10.6998	32.60825
baths	13995.46	3293.379	4.25	0.000	7515.581	20475.34
y81	35156.93	3039.97	11.56	0.000	29175.65	41138.22
_cons	-17094.5	9276.747	-1.84	0.066	-35346.95	1157.941

TEST FOR MULTICOLLINEARITY

Multicollinearity in regression occurs when two or more independent variables are highly correlated to each other, such that they do not provide unique information in the regression model.

To detect multicollinearity we use a metric known as the variance inflation factor (VIF), which measures the correlation and strength of correlation between the independent variables in a regression model.

$$VIF_i = 1 / 1 - R_i^2$$

where R_i^2 is the coefficient of determination of variable.

Since the VIF value for the independent variable is less than 5, there is very less multicollinearity in the model.

. vif

Variable	VIF	1/VIF
baths	2.74	0.364573
area	2.10	0.475366
rooms	1.71	0.584644
age	1.31	0.766010
y81	1.08	0.925662
nbh	1.03	0.975354
Mean VIF	1.66	

. pwcorr

	age	nbh	price	rooms	area	baths	y81
age	1.0000						
nbh	0.0748	1.0000					
price	-0.3319	-0.2159	1.0000				
rooms	-0.0512	-0.0645	0.4431	1.0000			
area	-0.0454	-0.0628	0.6453	0.5341	1.0000		
baths	-0.3569	-0.1176	0.6259	0.6038	0.6628	1.0000	
y81	-0.1104	-0.1005	0.5066	0.0058	0.1733	0.0471	1.0000

TEST FOR HETEROSCEDASTICITY

- If the error terms do not have constant variance, they are said to be heteroscedastic. The term means “differing variance”.
- Using the Breusch-Pagan / Cook-Weisberg test:

  Null hypothesis: Ho = equal variances for error

  Alt hypothesis: H1 = error variances changes as predicted values of Y increase
- The p-value that corresponds to the Chi-Square test statistic (89.20). In this case, it is 0.0000. Since this value is less than 0.05, we can reject the null hypothesis and conclude that heteroscedasticity is present in the data.
- Prob > F: This is the p-value that corresponds to the F test statistic. In this case, it is 0.0000. Since this value is less than 0.05, we can reject the null hypothesis and conclude that Strong heteroscedasticity is present in the data.

```
. imtest, white

White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(26)      =     89.20
         Prob > chi2   =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	89.20	26	0.0000
Skewness	9.82	6	0.1326
Kurtosis	-230477.82	1	1.0000
Total	-230378.80	33	1.0000

```
. hettest, rhs fstat

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: age nbh rooms area baths y81

         F(6 , 314)    =      7.68
         Prob > F      =    0.0000
```

RESOLVING HETEROSCEDASTICITY

The two main consequences of heteroskedasticity are;

1) ordinary least squares no longer produces the best estimators
2) standard errors computed using least squares can be incorrect and misleading.

Robust Standard Errors: As you can see, the Robust Standard Errors are very different from the Standard Errors drawn initally.

To remove heteroscedasticty we have taken robust standard error instead of normal standard error.

```
. reg price age nbh rooms area baths y81, vce(robust)

Linear regression                               Number of obs =       321
                                                F(  6,   314) =    116.74
                                                Prob > F      =    0.0000
                                                R-squared     =    0.7043
                                                Root MSE      =     23727
```

price	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-225.5894	54.6436	-4.13	0.000	-333.1033	-118.0755
nbh	-2109.249	618.7794	-3.41	0.001	-3326.727	-891.7707
rooms	4251.945	1725.201	2.46	0.014	857.5288	7646.361
area	21.65402	5.567451	3.89	0.000	10.6998	32.60825
baths	13995.46	3293.379	4.25	0.000	7515.581	20475.34
y81	35156.93	3039.97	11.56	0.000	29175.65	41138.22
_cons	-17094.5	9276.747	-1.84	0.066	-35346.95	1157.941

LOG LINEAR MODEL

In this model, we have taken the dependent variable in the log form, while all other explanatory variables are in their original form.

Comparing the value of R2 in the linear-linear model and this model we found that there is a significant increase in the value of R2 in this model i.e. about 5 per cent.

But only with an increase in the R2 value, we cannot strictly say that this model is better than the previous model because both models are in different forms and their coefficient interpretation would be different.

. reg ln_price age nbh rooms area baths y81

Source	SS	df	MS
Model	47.1764805	6	7.86274675
Residual	14.2625049	314	.04542199
Total	61.4389853	320	.191996829

Number of obs	=	321
F(6, 314)	=	173.10
Prob > F	=	0.0000
R-squared	=	0.7679
Adj R-squared	=	0.7634
Root MSE	=	.21312

ln_price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0030769	.000418	-7.36	0.000	-.0038993	-.0022545
nbh	-.014027	.0055738	-2.52	0.012	-.0249936	-.0030603
rooms	.0722366	.0172898	4.18	0.000	.0382182	.1062551
area	.0001835	.0000249	7.38	0.000	.0001346	.0002324
baths	.1607944	.0256082	6.28	0.000	.1104091	.2111797
y81	.3646419	.0248937	14.65	0.000	.3156625	.4136214
_cons	10.06474	.0928063	108.45	0.000	9.88214	10.24734

JOINT TEST

Here we have performed the joint significant test using the f test. We assume that the effect of area of the house, age of the house and no. of rooms in the house is the same on the dependent variable i.e. log price of the house. This is our null hypothesis and we can see from the test result that even at a 1 per cent significance level we can reject our null hypothesis.

Hence, our assumption is wrong. All these three explanatory variables will have different impacts on the price of the house (dependent variable).

```
. reg ln_price age nbh rooms area baths y81

      Source |       SS           df       MS      Number of obs   =       321
-------------+----------------------------------   F(  6,    314)  =    173.10
       Model |  47.1764805         6  7.86274675   Prob > F        =    0.0000
    Residual |  14.2625049       314  .04542199    R-squared       =    0.7679
-------------+----------------------------------   Adj R-squared   =    0.7634
       Total |  61.4389853       320  .191996829   Root MSE        =    .21312

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0030769    .000418    -7.36   0.000    -.0038993   -.0022545
         nbh |   -.014027   .0055738    -2.52   0.012    -.0249936   -.0030603
       rooms |   .0722366   .0172898     4.18   0.000     .0382182    .1062551
        area |   .0001835   .0000249     7.38   0.000     .0001346    .0002324
       baths |   .1607944   .0256082     6.28   0.000     .1104091    .2111797
         y81 |   .3646419   .0248937    14.65   0.000     .3156625    .4136214
       _cons |   10.06474   .0928063   108.45   0.000      9.88214    10.24734
------------------------------------------------------------------------------

. test age=area=rooms

 ( 1)  age - area = 0
 ( 2)  age - rooms = 0

       F(  2,   314) =   34.13
            Prob > F =    0.0000

.
```

DUMMY INTERACTION

Here we have interaction of Y81 which is a dummy variable with a quantitative explanatory variable, area.

But as you can see in the statistical analysis, the p-value of this variable is not less than the significant level i.e. 5 per cent.

```
. gen dinterctn1= y81* area

. reg ln_price age nbh rooms dinterctn1 area baths y81
```

Source	SS	df	MS
Model	47.2587406	7	6.75124865
Residual	14.1802448	313	.045304296
Total	61.4389853	320	.191996829

Number of obs	=	321
F(7, 313)	=	149.02
Prob > F	=	0.0000
R-squared	=	0.7692
Adj R-squared	=	0.7640
Root MSE	=	.21285

ln_price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.0032003	.0004274	-7.49	0.000	-.0040413	-.0023594
nbh	-.013307	.0055921	-2.38	0.018	-.0243099	-.0023041
rooms	.0720169	.0172681	4.17	0.000	.0380406	.1059932
dinterctn1	-.0000483	.0000358	-1.35	0.179	-.0001187	.0000222
area	.000211	.0000322	6.56	0.000	.0001477	.0002743
baths	.1577636	.0256737	6.14	0.000	.1072488	.2082785
y81	.4657978	.0790796	5.89	0.000	.3102029	.6213927
_cons	10.01903	.0986983	101.51	0.000	9.824836	10.21323

THANK YOU!