Steve Rodriguez

Model on Oregon Housing Price

For my research project I decided to find out which determinants affect Oregon's housing market the most. I wanted to do this project because I plan to stay in Oregon and I am heading into my senior year so future job prospects and housing are very important to me as the next steps in my life. There are also talks in my family right now about selling our current house and moving closer to the city for an easier commute to school and work so it is very relevant to me right now. When I was doing research to try and figure out some determinants of housing prices, I ran into a couple of interesting articles about the housing prices in Turkey and Sydney. The article about housing prices in Turkey by author Yener Coskun concluded that there is evidence of housing rent, construction cost and real mortgage interest rate having an impact on housing prices. In the article about Sydney, they concluded that the key long determinants of housing prices are gross disposable income, housing supply, unemployment rate and gross domestic product. After I read these articles I figured that the same thing can be applied to Oregon. My dependent variable ended up being property value, in dollars, and my independent variable was household income, also in dollars, travel time to work, in minutes, and lot size, in acres. The literature led me to believe that housing properties were not as important and to focus on bigger picture variables. My goal is to be able to find out what the biggest factor is in Oregon's housing prices so I have a better idea for my future and for anyone in the class to learn about it also considering they may also seek a house here.

I was able to get my data from the Public Use Microdata Sample that easily compiles data from the census bureau. There I had to make choices about my data very carefully because I wanted all my data to come from the same place to avoid any data errors and in doing so I had to settle for only a couple of variables that I deemed the most important. I ended up with over 29000 points of data which I was very happy about. However, when looking through the raw data I did find some problems. For the property value variable, the census bureau gives a house a “-1” dollar value if the house was vacant or not available for appraisal at that time. There were roughly 1300 of these values in my data set and theoretically I couldn’t allow them in my regression model because it doesn’t make sense to have a house have a property value of “-1” dollars. While looking through the data, I also found it very interesting to see so many “0”s in the travel time to work variable. I did end up keeping all those because that simply implies that those people work from home and I didnt have a theoretical reason to omit them. I ended up with 28190 observations with 28186 degrees of freedom.

My predictions for my data were very straight forward. I believed that household income would have a positive sign attached to it in my equation because for every dollar increase in household income I would expect property value to go up as theoretically I would expect people to want to have a nicer house as their income grew. For travel time to work I expected a negative sign because for every one more minute it takes to get to work, it would mean the house price would lower as it indicates a poor location and theoretically a house closer to the city with more jobs close by would see it’s price go up. Lot size I expected it to be positive, especially when it was in magnitudes such as acres, because the difference between a 1 and 2 acre size house is massive and more land bought means a higher price.

I then decided to try and do the expected estimates to see if I was right and I came up really rough. I followed the steps as best as I could and I believe I had some trouble because of how many values I had total and I may have messed up when inputting some commands into the google sheets. My expected results are in the appendix and the formula I used is also there. I had trouble with this formula in the past and I may have missed something, although I did try to double check it, the large data set may have been a problem and led me to results I didn't anticipate. I don't believe I have poor data as it is from the census and I already went through it trying to trim anything theoretically impossible, so the next thing I decided to do was a data dump to look at the sample statistics to see if I see anything wildly wrong. While looking at it, nothing stood out to me immediately other than the very large max in the travel time to work. That made me question if someone filled out their answer wrong because a 140+ minute commute to work makes me wonder if they counted the return back home in their calculation. I went ahead with my standard regression model because I wasn't sure what else to do with my data and I wanted to see how it was shaping up. As seen in the appendix I do believe I got the results I wanted. Here I saw coefficients and signs that I was expecting which makes me doubt the way I did it in the sheets. As you can further see however, the R squared and adjusted R square were all very low which makes me question if the variable I ended up choosing really did correlate well with the property value. In the regression model when we look at the P value for all the t-tests, we can see that there is evidence that all my variables have an impact on property value so that was good too see, the T values were also fairly high. The F test also passed and we are able to reject the null hypothesis and conclude that there is a difference between coefficients. Next thing I did was to check for the most important and concerning assumptions, multicollinearity, serial correlation, and heteroskedasticity.

For multicollinearity, I found it very easy to not find a problem as the T values were very high, the correlation table and VIF didn't show a problem so I concluded that there was no problem with multicollinearity. However, just for fun and to check off the assumption I decided to check if there was a problem with serial correlation, even though this wasn't a time series data. Unfortunately my data did show some problem with serial correlation, as seen in the appendix, and I believe the root cause was an omitted variable. Even though lagging the error doesn't really make sense in this context as it is taking the difference between the previous data point and the next one, which doesn't make sense in this context. I do believe I had the correct functional form and none of the variables that I had could be considered irrelevant so I concluded it was an omitted variable. I also wouldn't be able to do any of the remedies to serial correlation so I wasn't able to do things such as the durbin watson test. The omitted variable problem can also be seen when I did the heteroskedasticity assumption check. In the appendix you can see that it is easily detectable that heteroskedasticity is a problem through the park test and residual plots. I then tried to fix the problem with a log log model, but I still failed the park test. I then tried the weighted least squares, and robust command to try and fix the standard errors as I knew the coefficients were unbiased but it still didn't seem to help. I believe this again is another problem arising from an omitted variable.

When I look at the total sum of squares which is the error explained by my model and the residual sum of squares which is the error not explained by my model, I can very clearly see that there is much more unexplained than explained. I believe that with serial correlation and heteroskedasticity facing problems with omitted variables and the R squared being so low, the main remedy would be to go back and gather more data such as bedrooms in a house or even unemployment rate.

Bibliography

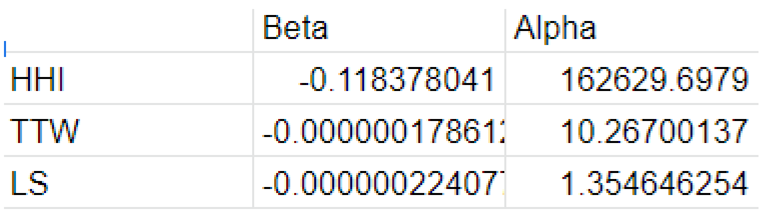
Coskun, Yener, et al. “Housing Price Dynamics and Bubble Risk: The Case of Turkey.” *Housing Studies*, vol. 35, no. 1, Jan. 2020, pp. 50–86. *EBSCOhost*, doi:http://www-tandfonline-com.proxy.lib.pdx.edu/loi/chos20.

Al-Masum, Md Abdullah, and Chyi Lin Lee. “Modelling Housing Prices and Market Fundamentals: Evidence from the Sydney Housing Market.” *International Journal of Housing Markets and Analysis*, vol. 12, no. 4, 2019, pp. 746–762. *EBSCOhost*, search.ebscohost.com/login.aspx?direct=true&db=ecn&AN=1802978&site=ehost-live.

Appendix

Formula for expected coefficients and results

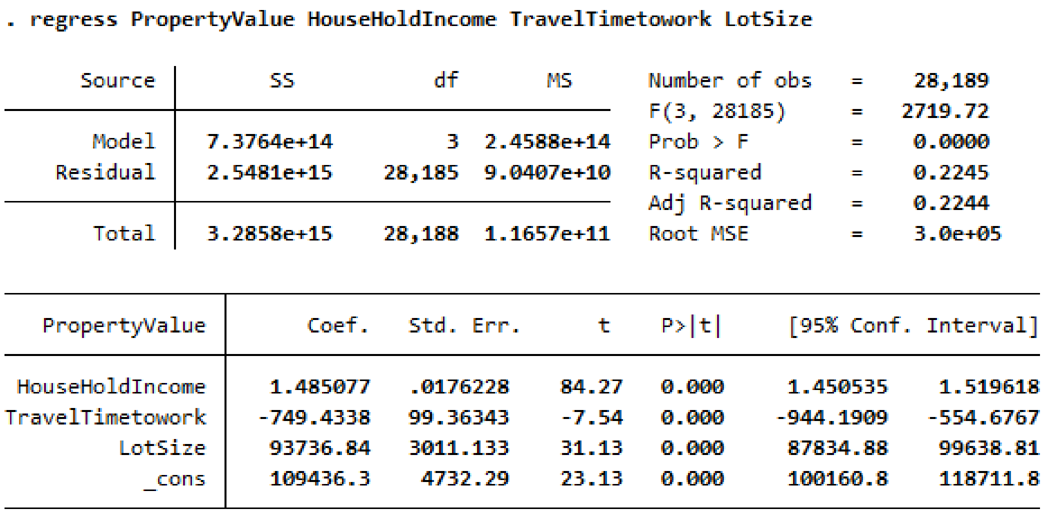
$$\beta = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2}$$



	Beta	Alpha
HHI	-0.118378041	162629.6979
TTW	-0.00000017861	10.26700137
LS	-0.000000224077	1.354646254

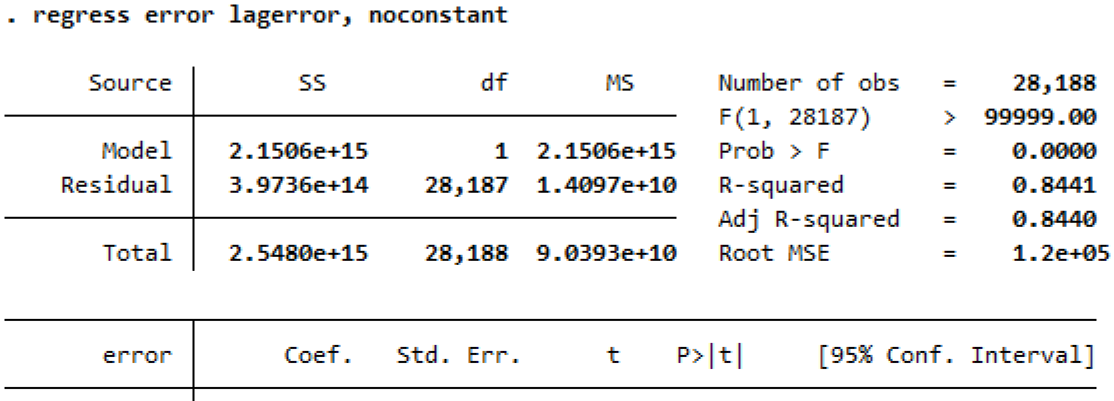
$$\alpha = \overline{Y} - \beta\overline{X}$$

Standard Regression



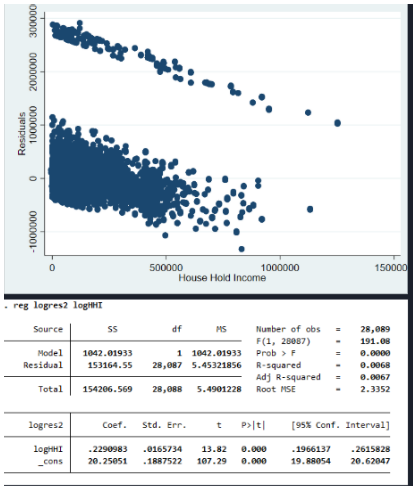
```
. regress PropertyValue HouseHoldIncome TravelTimetowork LotSize

      Source |       SS           df       MS      Number of obs   =    28,189
-------------+----------------------------------   F(3, 28185)     =   2719.72
       Model |  7.3764e+14         3  2.4588e+14   Prob > F        =    0.0000
    Residual |  2.5481e+15    28,185  9.0407e+10   R-squared       =    0.2245
-------------+----------------------------------   Adj R-squared   =    0.2244
       Total |  3.2858e+15    28,188  1.1657e+11   Root MSE        =    3.0e+05

------------------------------------------------------------------------------
PropertyValue |     Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
HouseHoldIncome |  1.485077   .0176228    84.27   0.000     1.450535    1.519618
TravelTimetowork | -749.4338   99.36343    -7.54   0.000    -944.1909   -554.6767
      LotSize |   93736.84   3011.133    31.13   0.000     87834.88    99638.81
        _cons |   109436.3    4732.29    23.13   0.000     100160.8    118711.8
------------------------------------------------------------------------------
```

Serial correlation test

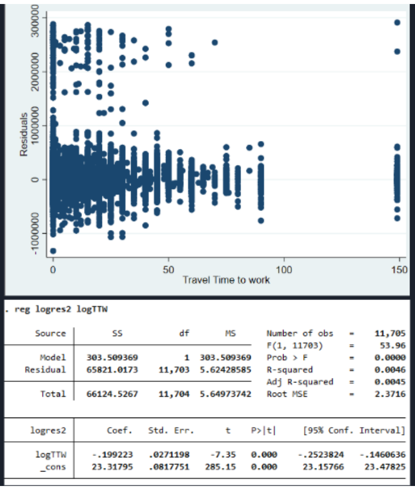


```
. regress error lagerror, noconstant

      Source |       SS           df       MS      Number of obs   =    28,188
-------------+----------------------------------   F(1, 28187)     >  99999.00
       Model |  2.1506e+15         1  2.1506e+15   Prob > F        =    0.0000
    Residual |  3.9736e+14    28,187  1.4097e+10   R-squared       =    0.8441
-------------+----------------------------------   Adj R-squared   =    0.8440
       Total |  2.5480e+15    28,188  9.0393e+10   Root MSE        =    1.2e+05

------------------------------------------------------------------------------
       error |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
```

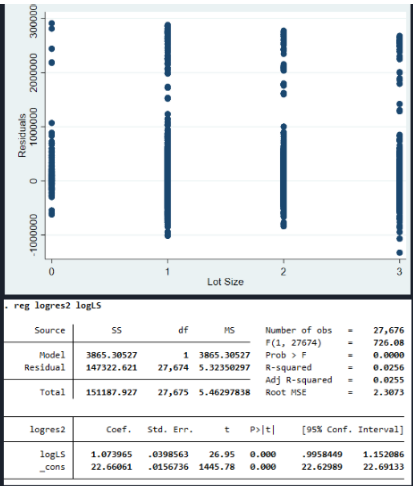
Heteroskedasticity test



```
. reg logres2 logHHI

      Source |       SS           df       MS      Number of obs   =    28,089
-------------+----------------------------------   F(1, 28087)     =    191.08
       Model |  1042.01933         1  1042.01933   Prob > F        =    0.0000
    Residual |   153164.55    28,087  5.45321856   R-squared       =    0.0068
-------------+----------------------------------   Adj R-squared   =    0.0067
       Total |  154206.569    28,088   5.4901228   Root MSE        =    2.3352

------------------------------------------------------------------------------
     logres2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logHHI |   .2290983   .0165734    13.82   0.000     .1966137    .2615828
       _cons |   20.25051   .1887522   107.29   0.000     19.88054    20.62047
------------------------------------------------------------------------------
```

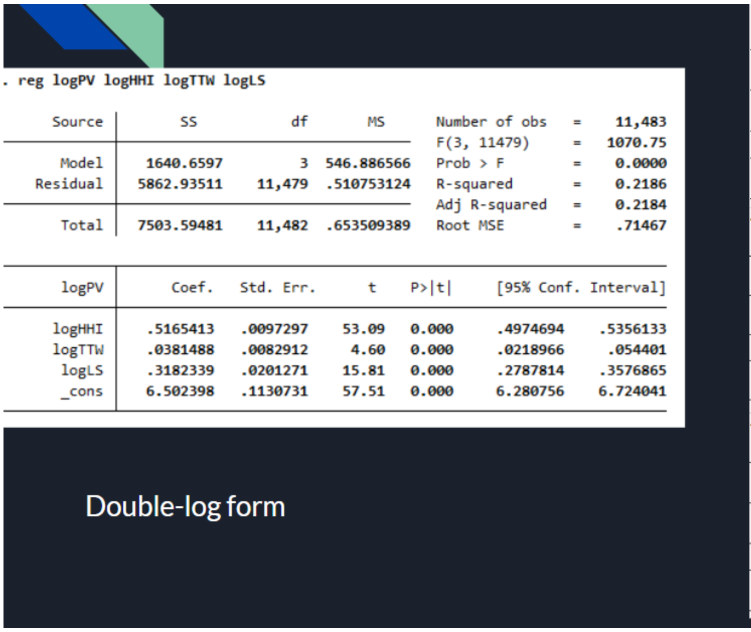


```
. reg logres2 logTTW

      Source |       SS           df       MS      Number of obs   =    11,705
-------------+----------------------------------   F(1, 11703)     =     53.96
       Model |  303.509369         1  303.509369   Prob > F        =    0.0000
    Residual |  65821.0173    11,703  5.62428585   R-squared       =    0.0046
-------------+----------------------------------   Adj R-squared   =    0.0045
       Total |  66124.5267    11,704  5.64973742   Root MSE        =    2.3716

------------------------------------------------------------------------------
     logres2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logTTW |   -.199223   .0271198    -7.35   0.000    -.2523824   -.1460636
       _cons |   23.31795   .0817751   285.15   0.000     23.15766    23.47825
------------------------------------------------------------------------------
```



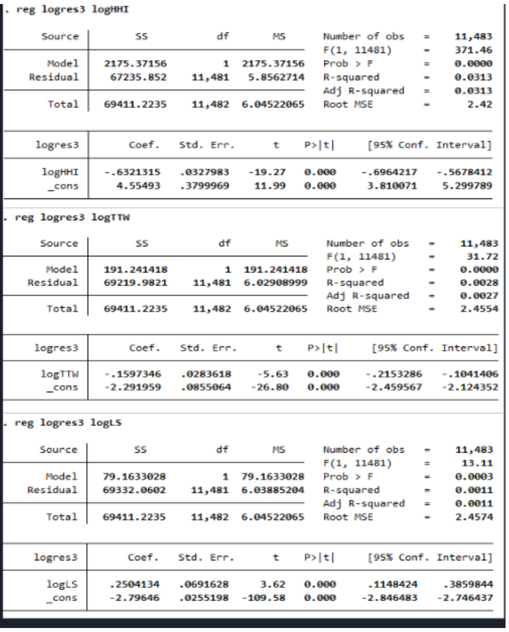
```
. reg logres2 logLS

      Source |       SS           df       MS      Number of obs   =    27,676
-------------+----------------------------------   F(1, 27674)     =    726.08
       Model |  3865.30527         1  3865.30527   Prob > F        =    0.0000
    Residual |  147322.621    27,674  5.32350297   R-squared       =    0.0256
-------------+----------------------------------   Adj R-squared   =    0.0255
       Total |  151187.927    27,675  5.46297838   Root MSE        =    2.3073

------------------------------------------------------------------------------
     logres2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       logLS |   1.073965   .0398563    26.95   0.000     .9958449    1.152086
       _cons |   22.66061   .0156736  1445.78   0.000     22.62989    22.69133
------------------------------------------------------------------------------
```

Heteroskedasticity remedy



```
. reg logPV logHHI logTTW logLS

      Source |       SS           df       MS      Number of obs   =    11,483
-------------+----------------------------------   F(3, 11479)     =   1070.75
       Model |   1640.6597         3  546.886566   Prob > F        =    0.0000
    Residual |  5862.93511    11,479  .510753124   R-squared       =    0.2186
-------------+----------------------------------   Adj R-squared   =    0.2184
       Total |  7503.59481    11,482  .653509389   Root MSE        =    .71467

------------------------------------------------------------------------------
       logPV |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logHHI |   .5165413   .0097297    53.09   0.000     .4974694    .5356133
      logTTW |   .0381488   .0082912     4.60   0.000     .0218966     .054401
       logLS |   .3182339   .0201271    15.81   0.000     .2787814    .3576865
       _cons |   6.502398   .1130731    57.51   0.000     6.280756    6.724041
------------------------------------------------------------------------------
```

Double-log form



```
. reg logres3 logHHI

      Source |       SS           df       MS      Number of obs   =    11,483
-------------+----------------------------------   F(1, 11481)     =    371.46
       Model |  2175.37156         1  2175.37156   Prob > F        =    0.0000
    Residual |   67235.852    11,481   5.8562714   R-squared       =    0.0313
-------------+----------------------------------   Adj R-squared   =    0.0313
       Total |  69411.2235    11,482  6.04522065   Root MSE        =      2.42

------------------------------------------------------------------------------
     logres3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logHHI |  -.6321315   .0327983   -19.27   0.000    -.6964217   -.5678412
       _cons |    4.55493   .3799969    11.99   0.000     3.810071    5.299789
------------------------------------------------------------------------------

. reg logres3 logTTW

      Source |       SS           df       MS      Number of obs   =    11,483
-------------+----------------------------------   F(1, 11481)     =     31.72
       Model |  191.241418         1  191.241418   Prob > F        =    0.0000
    Residual |  69219.9821    11,481  6.02908999   R-squared       =    0.0028
-------------+----------------------------------   Adj R-squared   =    0.0027
       Total |  69411.2235    11,482  6.04522065   Root MSE        =    2.4554

------------------------------------------------------------------------------
     logres3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      logTTW |  -.1597346   .0283618    -5.63   0.000    -.2153286   -.1041406
       _cons |  -2.291959   .0855064   -26.80   0.000    -2.459567   -2.124352
------------------------------------------------------------------------------

. reg logres3 logLS

      Source |       SS           df       MS      Number of obs   =    11,483
-------------+----------------------------------   F(1, 11481)     =     13.11
       Model |  79.1633028         1  79.1633028   Prob > F        =    0.0003
    Residual |  69332.0602    11,481  6.03885204   R-squared       =    0.0011
-------------+----------------------------------   Adj R-squared   =    0.0011
       Total |  69411.2235    11,482  6.04522065   Root MSE        =    2.4574

------------------------------------------------------------------------------
     logres3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       logLS |   .2504134   .0691628     3.62   0.000     .1148424    .3859844
       _cons |   -2.79646   .0255198  -109.58   0.000    -2.846483   -2.746437
------------------------------------------------------------------------------
```