Data Exercise for Homework 1 of Data Analysis 2

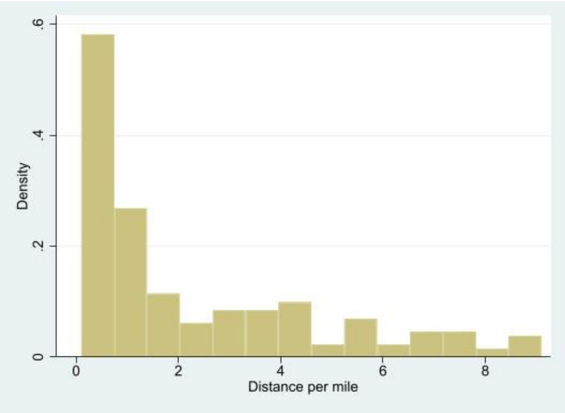
Part A: Level-Level Model

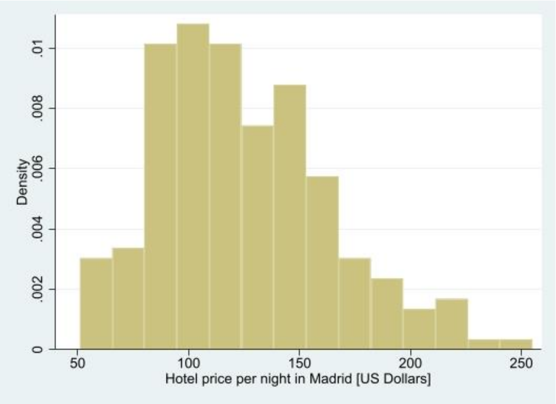
```
      Source |       SS           df       MS      Number of obs   =       203
-------------+----------------------------------   F(1, 201)       =     16.80
       Model |  23741.9512         1  23741.9512   Prob > F        =    0.0001
    Residual |  283996.187       201  1412.91635   R-squared       =    0.0771
-------------+----------------------------------   Adj R-squared   =    0.0726
       Total |  307738.138       202  1523.45613   Root MSE        =    37.589

------------------------------------------------------------------------------
       price | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    distance |  -4.534921   1.106292    -4.10   0.000    -6.716347   -2.353494
       _cons |   135.4811    3.65059    37.11   0.000     128.2828    142.6795
------------------------------------------------------------------------------
```

For this data exercise I have created a simple linear regression for the city of Madrid. Utilizing the provided data sets, I have selected the number of stars (3 and 4), weekend (0), and month (11) for the city of Madrid. From this table I can see that the intercept is $135.48 and the slope is -4.5. This signifies that when distance is 0, the average price per night is approximately $135.48. As we go 1 mile from the center, prices are lower on average by $4.53 dollars.

In class, our intercept and slope for Vienna was approximately $132 and -14 respectively. This means that at the city center, on average, hotel prices were $132 and at each unit (which may have been miles in this example) away from the center, prices are on average $14 lower.





Data Exercise for Homework 1 of Data Analysis 2

Part B: Log-Level Model, Natural Log of Y

```
      Source |       SS           df       MS      Number of obs   =       203
-------------+----------------------------------   F(1, 201)       =     20.35
       Model |  1.82343727         1  1.82343727   Prob > F        =    0.0000
    Residual |  18.0129673       201  .089616753   R-squared       =    0.0919
-------------+----------------------------------   Adj R-squared   =    0.0874
       Total |  19.8364045       202  .098200022   Root MSE        =    .29936

------------------------------------------------------------------------------
    logprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
    distance |  -.0397427   .0088106    -4.51   0.000    -.0571157   -.0223696
       _cons |    4.87195   .0290736   167.57   0.000     4.814622    4.929279
------------------------------------------------------------------------------
```

In this question, I have taken the natural log of price, y. With taking the natural log of price, we can interpret this as, when we go 1 mile away from the center prices are (-.03 * 100) -3.896 percent lower. The average of the logprice of hotels at the city center is $129.57 (e^4.87195-1).

Part C: Log-Log Model, Natural Log of x,y

```
      Source |       SS           df       MS      Number of obs   =       203
-------------+----------------------------------   F(1, 201)       =     12.90
       Model |  1.19665011         1  1.19665011   Prob > F        =    0.0004
    Residual |  18.6397544       201  .092735097   R-squared       =    0.0603
-------------+----------------------------------   Adj R-squared   =    0.0557
       Total |  19.8364045       202  .098200022   Root MSE        =    .30452

------------------------------------------------------------------------------
    logprice | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 logdistance |  -.0611711   .0170288    -3.59   0.000    -.0947491    -.027593
       _cons |   4.791114   .0215471   222.36   0.000     4.748627    4.833602
------------------------------------------------------------------------------
```

For this question, I have taken the natural log for both variables x,y, or a log-log model. With our new calculation using the log-log model, our alpha is 4.79 and our log distance is -0.6. We interpret this as: as we go one mile away from the center, the price of hotels is approximately 6 percent lower with the logprice on average being $119.43 (e^4.791114-1) at the city center (when logdistance is 0). Hotels that are one percent away from the center are on average 6% lower.

We can compare the R-squared for the Log-Log and Log-Level models in this scenario and see that the Log-Level model has an R-squared of .09 and the Log-Log model has an R-squared model of .06. R-squared highlights the variance in y. From this we can say that 9% and 6%, respectively, of overall variation in hotel prices is explained by the linear regression with distance to the city center. This leaves 91% and 94% unexplained, respectively. There is greater variance in y in the Log-Level model though both have a relatively low variation.

Do file on Word

```
*Locate your path file, where the data is.
global path "C:\Users\Mosby_Victoria\Desktop"
```

Data Exercise for Homework 1 of Data Analysis 2

```
*Open data set.
Use "$path/hotels-europe_features.dta", clear

*Before merging, sort the hotel_id variable in numerical order.
sort hotel_id

*Merge the price data set with the features data set using the hotel_id variable.
merge 1:m hotel_id using "$path/hotels-europe_price.dta"
tab _m
drop _m

*Follow instructions by eliminating unneccessary data for the regression analysis.
keep if stars>3
drop if stars<4
keep if accommodation_type =="Hotel"
keep if city_actual == "Madrid"
keep if year == 2017
keep if month == 11
keep if weekend == 0

*Descriptive statistics
summarize price, det
summarize distance, det

*Graphs
hist price
hist distance

*Drop outliers
drop if price> 600
drop if price>300

*Label the price and distance variables
label variable price "Hotel price per night in Madrid (US Dollars)"
label variable distance "Distance per mile"

*Linear Regression
regress price distance
```

```
Source      SS          df          MS      Number of obs   =       203
                                    F(1, 201)       =       16.80
Model       23741.9512              1           23741.9512   Prob > F       =  0.0001
Residual    283996.187              201         1412.91635   R-squared      = 0.0771
                                    Adj R-squared   =    0.0726
Total       307738.138              202         1523.45613   Root MSE       =37.589


price       Coefficient             Std. err.   t    P>t    [95% conf.              interval]

distance    -4.534921               1.106292 -4.10   0.000   -6.716347      -2.353494
_cons       135.4811 3.65059        37.11   0.000   128.2828      142.6795
```

```
*distance x, price y
*intercept:135 this means that when the distance is 0, the price is $135
*slope: -4.53, for every mile away from center, prices on average $4.53 lower

*Log-Level model
gen ln_price = ln(price)
reg ln_price distance
```

```
            Source      SS          df          MS      Number of obs   =       203
                                                F(1, 201)       =       20.35
            Model       1.82343727              1           1.82343727   Prob > F       =  0.0000
            Residual    18.0129673              201         .089616753   R-squared      = 0.0919
                                                Adj R-squared   =    0.0874
            Total       19.8364045              202         .098200022   Root MSE       =.29936


            ln_price    Coefficient             Std. err.   t    P>t    [95% conf.              interval]

            distance    -.0397427               .0088106 -4.51   0.000   -.0571157      -.0223696
            _cons       4.87195    .0290736 167.57   0.000    4.814622      4.929279
```

Data Exercise for Homework 1 of Data Analysis 2

```
*distance, lnprice
*as we go one mile away from center, prices are on average 3.97% lower (-.0397*100)
*lnprice shows percentage, better approximation to average slope

*Log-log model
gen ln_distance = ln(distance)
regress ln_price ln_distance

Source    SS          df          MS      Number of obs   =       203
                                  F(1, 201)       =        12.90
Model     1.19665011              1           1.19665011   Prob > F        =   0.0004
Residual  18.6397544              201         .092735097   R-squared       = 0.0603
                                  Adj R-squared   =    0.0557
Total     19.8364045              202         .098200022   Root MSE        =.30452


ln_price   Coefficient            Std. err.   t    P>t     [95% conf.              interval]

ln_distance            -.0611711              .0170288 -3.59   0.000    -.0947491        -.027593
_cons      4.791114 .0215471 222.36   0.000     4.748627      4.833602

*ln_distance, ln_price
*e^4.791114-1 119$ average ln_price
*as we move one percent away from center, prices on average are 6% lower

save as "C:\Users\Mosby_Victoria\Desktop\Data 2\Data HW 2 Regression Analysis Madrid.dta"
```