Practical 7 Regression

This week, we will continue to use the dataset "nssec_houseprice_airbnb.sav" from week 6, to build up various regression models on Airbnb Listing Properties' prices and consolidate your capability to interpret the outcomes.

This Week's Overview:

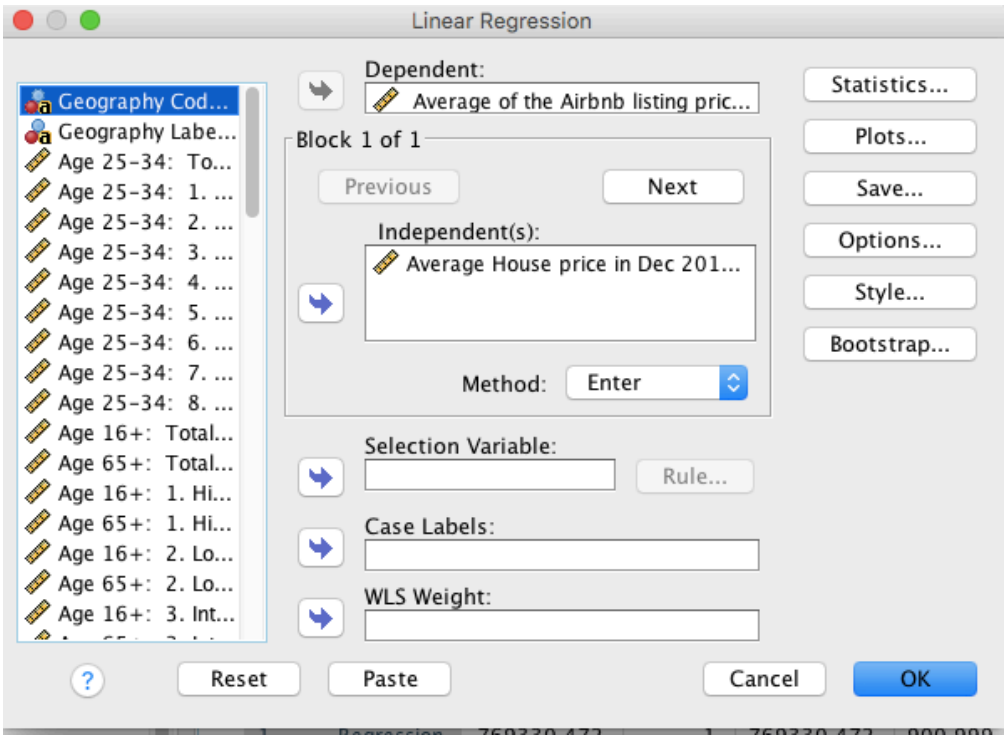
- We're going to build up simple linear regression (OLS) model.
- We're going build up multiple linear regression model.
- We're going to build up automatic linear regression model.
- We're also going to conduct regression diagnostics on
    - normality
    - model specification
    - independence
    - multicollinearity
    - identifying the unusual and influential data.

Simple Linear Regression

Let's create a simple linear regression that uses only one variable, House Price, for Airbnb Price. We are wondering how much of the average Airbnb property's rental price we can predict based solely on the regional average house price. To obtain this, we could select from the menu that:

```
Analyse -> Regression -> Linear
```

To call up the dialog box shown below, and please make sure that you turn on 95% Confidence Intervals in the Options.



It will be great to run the regressions in the following order:

1) Airbnb Price against House Price using the full data set
2) Airbnb Price against House Price using a random sample of 98 records. So

what do you think is the dependent variable? Add this to the appropriate box;

what do you think is the independent variable? Add this to the appropriate box.

The output of this most basic case produces a summary chart showing ***R***, ***R-square***, and the ***Standard error*** of the prediction; an ***ANOVA chart***; and a chart providing ***statistics on model coefficients***:

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Average House price in Dec 2017[b]	.	Enter

a. Dependent Variable: Average of the Airbnb listing prices

b. All requested variables entered.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.694[a]	.481	.481	29.2209557

a. Predictors: (Constant), Average House price in Dec 2017

b. Dependent Variable: Average of the Airbnb listing prices

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	769330.472	1	769330.472	900.999	.000[b]
	Residual	829956.051	972	853.864		
	Total	1599286.52	973			

a. Dependent Variable: Average of the Airbnb listing prices

b. Predictors: (Constant), Average House price in Dec 2017

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	34.726	1.733		20.035	.000	31.325	38.127
	Average House price in Dec 2017	7.236E-5	.000	.694	30.017	.000	.000	.000

a. Dependent Variable: Average of the Airbnb listing prices

Track your results in the table below:

Model Number	R^2	Intercept (CIs)	House (CIs)	Near-Normal Residuals (Y/N)
1. Airbnb/House Full		/	/	
2. Airbnb/House Samp		/	/	

Q1. ***What is the effect of House Price on Airbnb Price? In other words, is there a positive or negative impact? How can you tell?***

Q2. What is the range (CIs) of the constant? In other words, what range does the model predict for the average price of Airbnb property where no houses for sale?

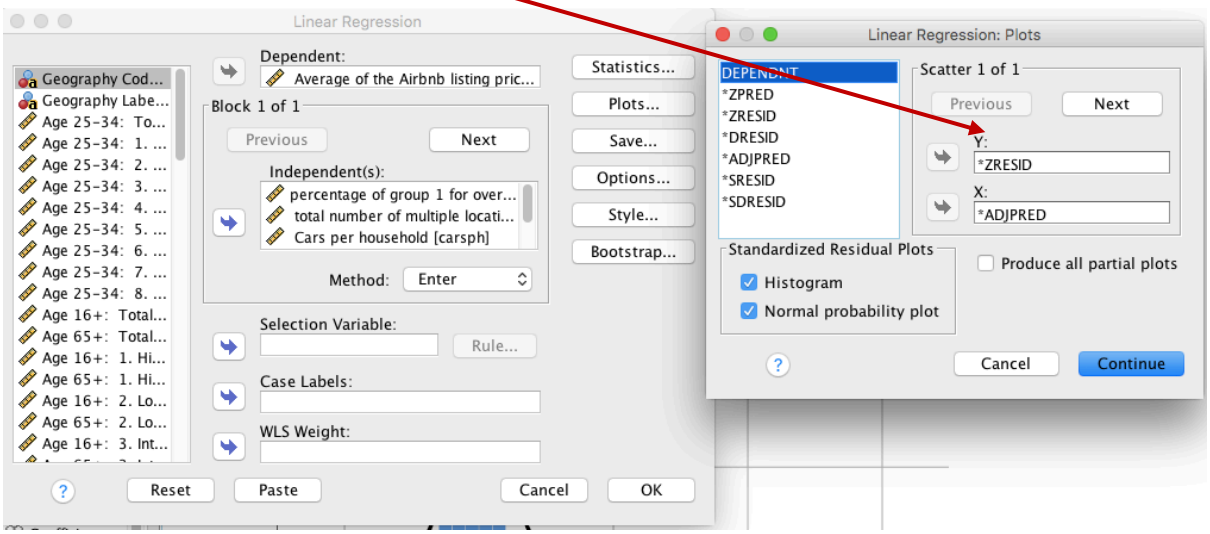
Q3. What effect does the smaller sample size have on your model, and which model offers a better prediction?

Multiple Linear Regression

If you'd like to dig into regression further, then UCLA has a nice set of web resources: http://stats.idre.ucla.edu/spss/webbooks/reg/chapter2/spss-webbooksregressionwith-spsschapter-2-regression-diagnostics/. There are other resources on SPSS provided by IDRE at www.ats.ucla.edu/stat/spss/output/reg_spss.htm and material on the residual plot can be found here: www.ats.ucla.edu/stat/spss/webbooks/reg/chapter1/spssreg1.htm.

Let's first try seeing what happens with a couple of variables and, in particular, let's see what happens when we use house price and multiple location listings:

- Run the regression with *Mean Airbnb listing price* as dependent variable, and for example choose, *percentage of Group 1 in all residents over 16*, *average house prices*, *multiple location Airbnb listings*, *average number of cars* in each household and *average household income* as the input variables.
- Check the options of "**statistics**" and "**plot**" by ticking the 95% confidence intervals and get **residuals plotted**.



Q4. How do you interpret the results of this regression?

Q5. Which parameter is more significant to your results?

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	48.876	3.373		14.490	.000	42.257	55.495
	percentage of group 1 for over 16	101.705	31.510	.169	3.228	.001	39.870	163.541
	total number of multiple location hosts listed airbnbs	.178	.020	.228	8.754	.000	.138	.218
	Cars per household	-20.104	3.333	-.166	-6.031	.000	-26.645	-13.562
	Total Mean Annual Household Income	3.757E-5	.000	.014	.218	.828	.000	.000
	Average House price in Dec 2017	4.564E-5	.000	.437	11.545	.000	.000	.000

a. Dependent Variable:
Average of the Airbnb listing prices

Q6. What is the effect on the y-intercept and how do you interpret this result?

Q7. What is the difference between the unstandardised and the standardised coefficients reported by SPSS?

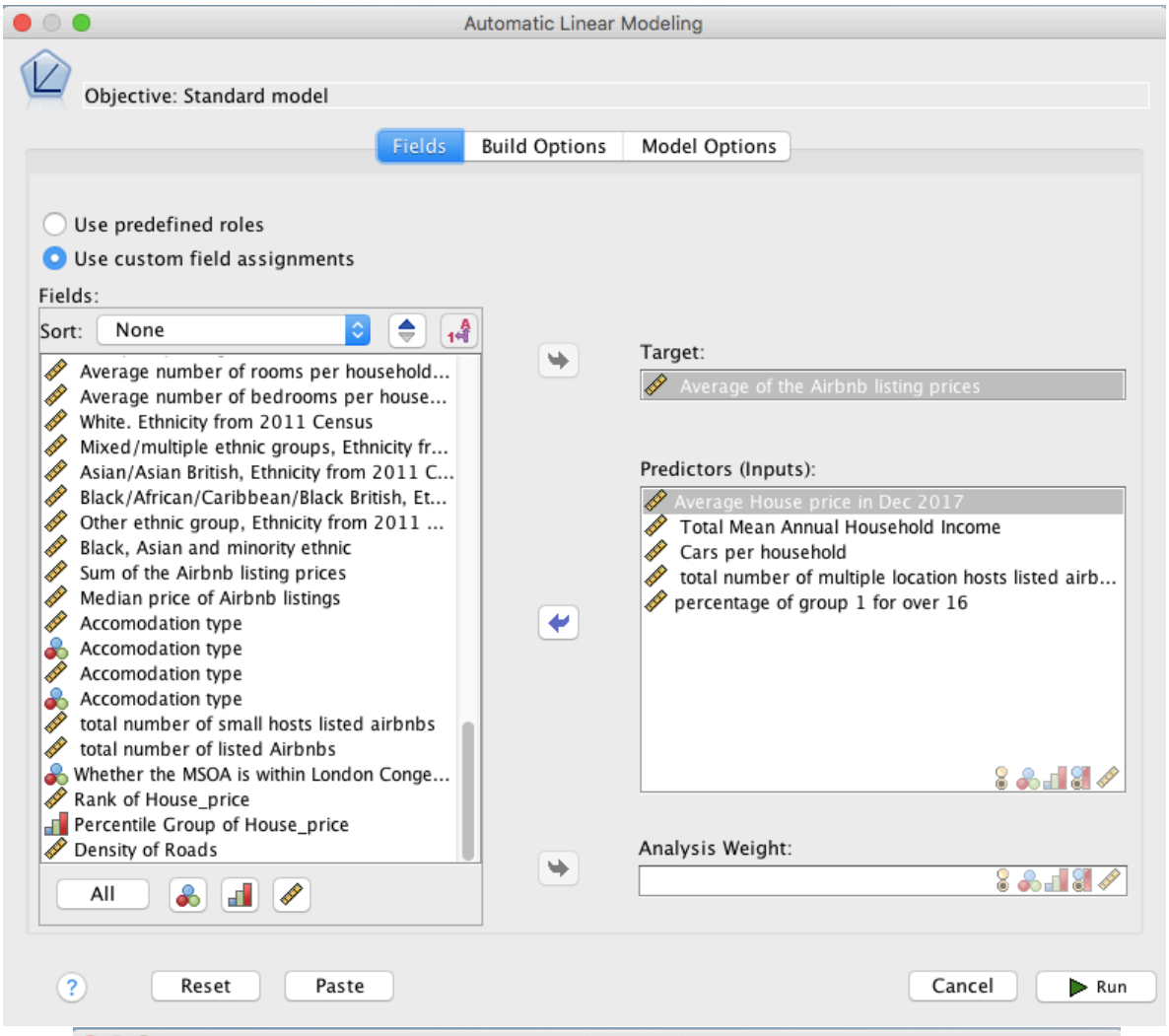
Q8. How do I interpret the residual plot?

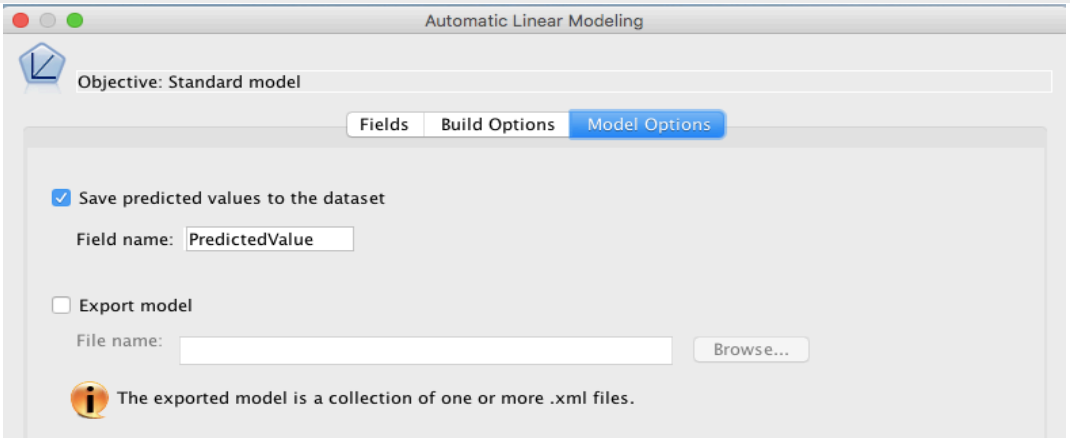
Automatic Linear Modelling (Optional, you may leave it later)

We can't properly add all of the other variables to a simple linear model, so we need to switch to '**Automatic Linear Modelling**' and it allows regression to automatically cope with a range of variables and residual types. This is the menu option immediately above "Linear".

`Analyze->Regression->Automatic Linear Modelling`

- Re-run the regression using:
  - Mean Airbnb listing price
  - percentage of Group 1 in all residents over 16
  - average house prices
  - multiple location Airbnb listings
  - average number of cars in each household
  - average household income
- Make sure that you specify `: Save predicted values` (under Model Options)



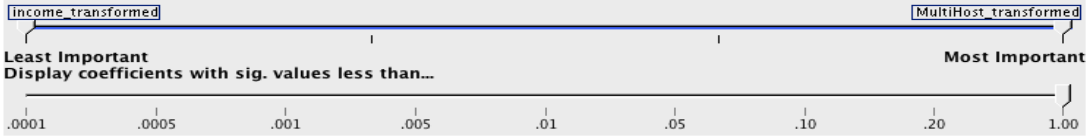


To access the ALM results you'll need to double-click the output in the Output Window, and some of the windows have scroll down options to display in various formats.

Coefficients

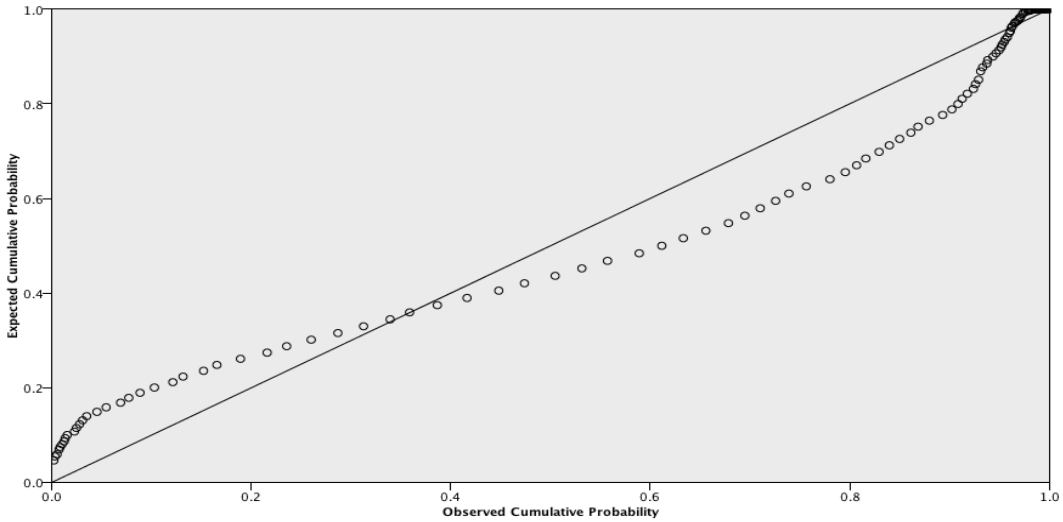
Target:
Average of the Airbnb listing prices

Model Term	Coefficient ▼	Std.Error	t	Sig.	95% Confidence Interval		Importance
					Lower	Upper	
Intercept	28.155	3.716	7.576	.000	20.862	35.448	
MultiHost_transformed	0.485	0.054	9.046	.000	0.380	0.591	0.429
House_price_transformed	0.000	0.000	9.004	.000	0.000	0.000	0.425
carsph_transformed	-14.237	3.753	-3.794	.000	-21.602	-6.872	0.075
income_transformed	0.000	0.000	3.659	.000	0.000	0.001	0.070



Residuals

Target:
Average of the Airbnb listing prices



The P-P plot of Studentized residuals compares the distribution of the residuals to a normal distribution.The diagonal line represents the normal distribution.The closer the observed cumulative probabilities of the residuals are to this line, the closer the distribution of the residuals is to the normal distribution.

Questions

1. *How do you interpret the results of this regression*?

2. *How much of property price have we 'explained' with the new model*?

3. *What is the most useful predictor of Airbnb listing price*?

4. *What is the least useful predictor of property price*?

5. *What is the y-intercept*?

Now let's add borough to the model and re-run the Automatic Linear Model.

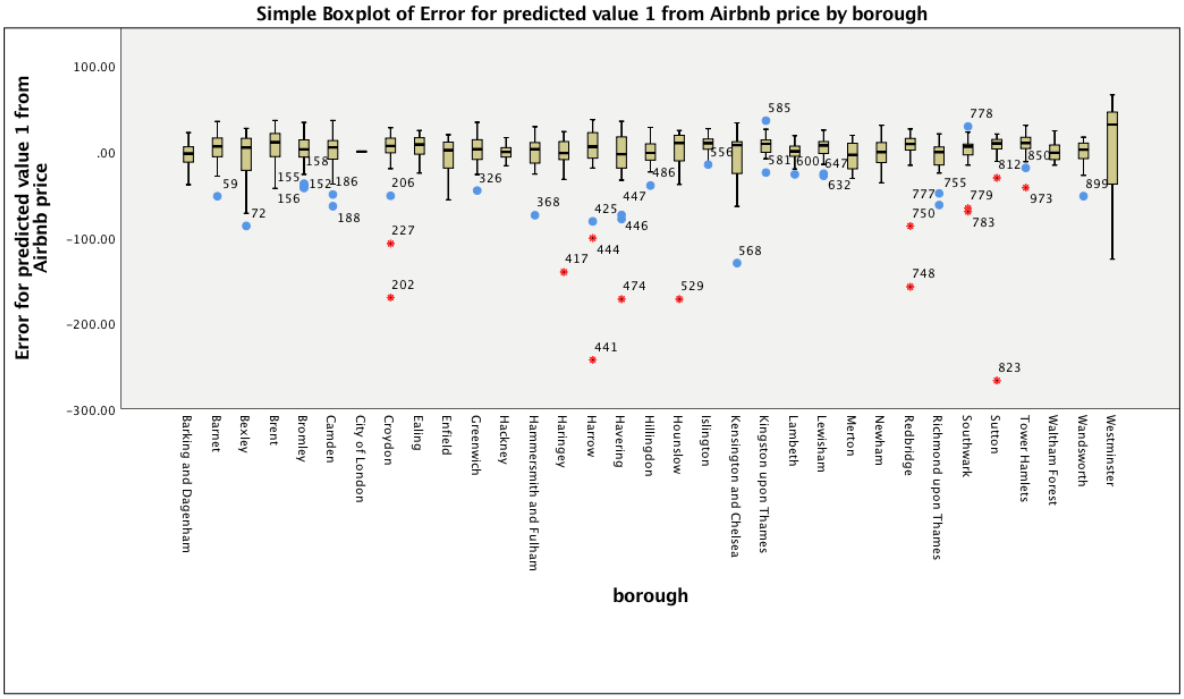
- Move the Borough ("borough") from the 'Fields' section to the 'Predictors' section.
- Run the model and, under '**Model Options**', save the predicted value to a new field called 'Predicted Value 1'.

6. *Has the model been noticeably improved and can you explain the outcomes from Borough_Transformed (broadly)*?

7. *Why do you think that the "borough" variable has a strong effect on the model*?

8. *Can you explain why the importance of other variables appears to have changed*?

9. *Explain the boxplot…* Try to figure out how I produced the boxplot for Airbnb Pricing to determine if there is evidence of bias in our model residuals that could be geographic or otherwise.

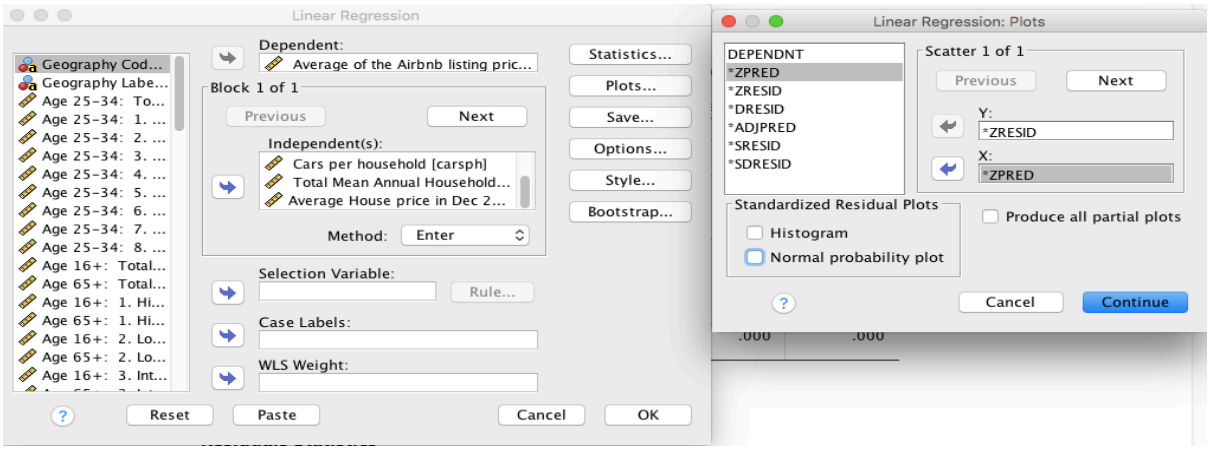


Regression Diagnostics—Mapping the residuals

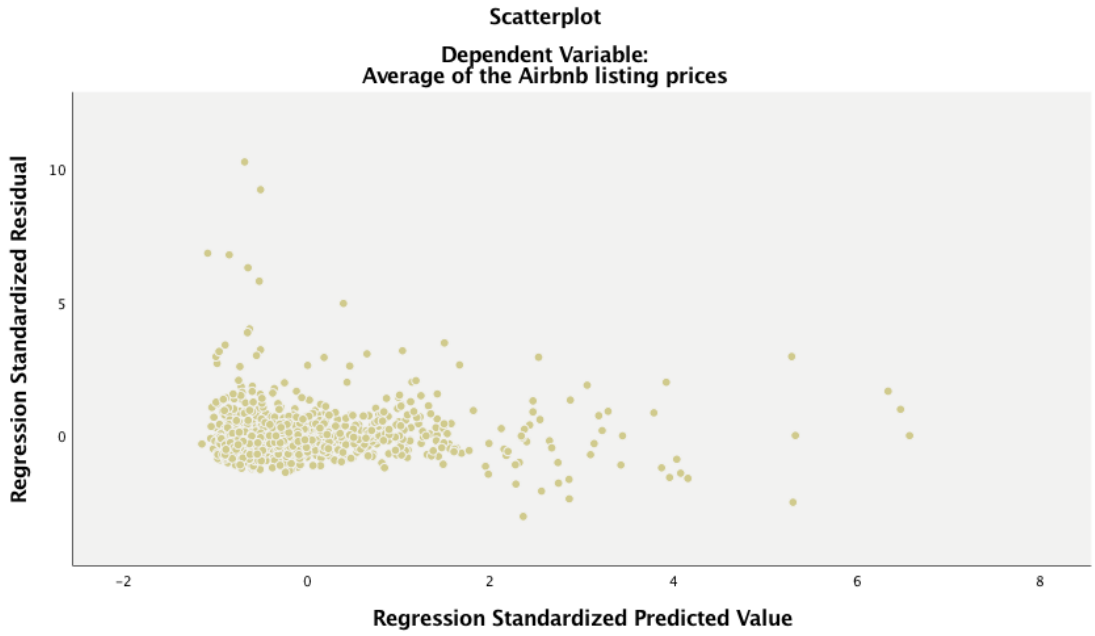
When we do linear regression, we assume that the relationship between the response variable and the predictors is linear, which could be visually reflected by a straight fit line to the data. However, the received estimated coefficients and standard errors might be biased due to failing to check the assumption of linear regression (e.g., you can get a significant effect when in fact there is none, or vice versa). In order to check whether the data meet the assumption, you need to check the residuals' normality, and you could resort to a bivariate plot for outcome

visualisation, which plot a scatterplot of the standardized predicted value against standardized residuals.

`Analyze – Regression – Linear`



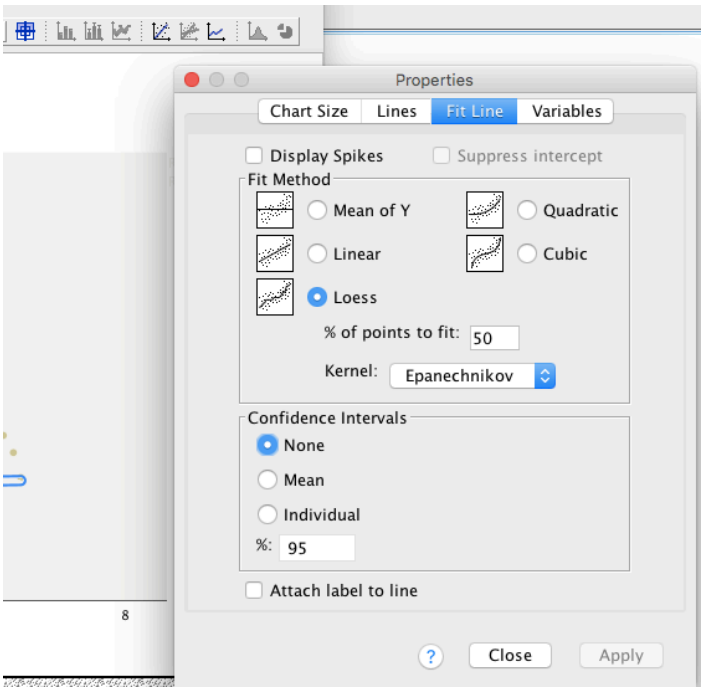
Similar setting with multiple linear regression but choose different variables to plot as shown above (ZRESID to the Y: field and *ZPRED to the X: field). You will receive a scatterplot of the standardized residuals with the standardized predicted values as below:



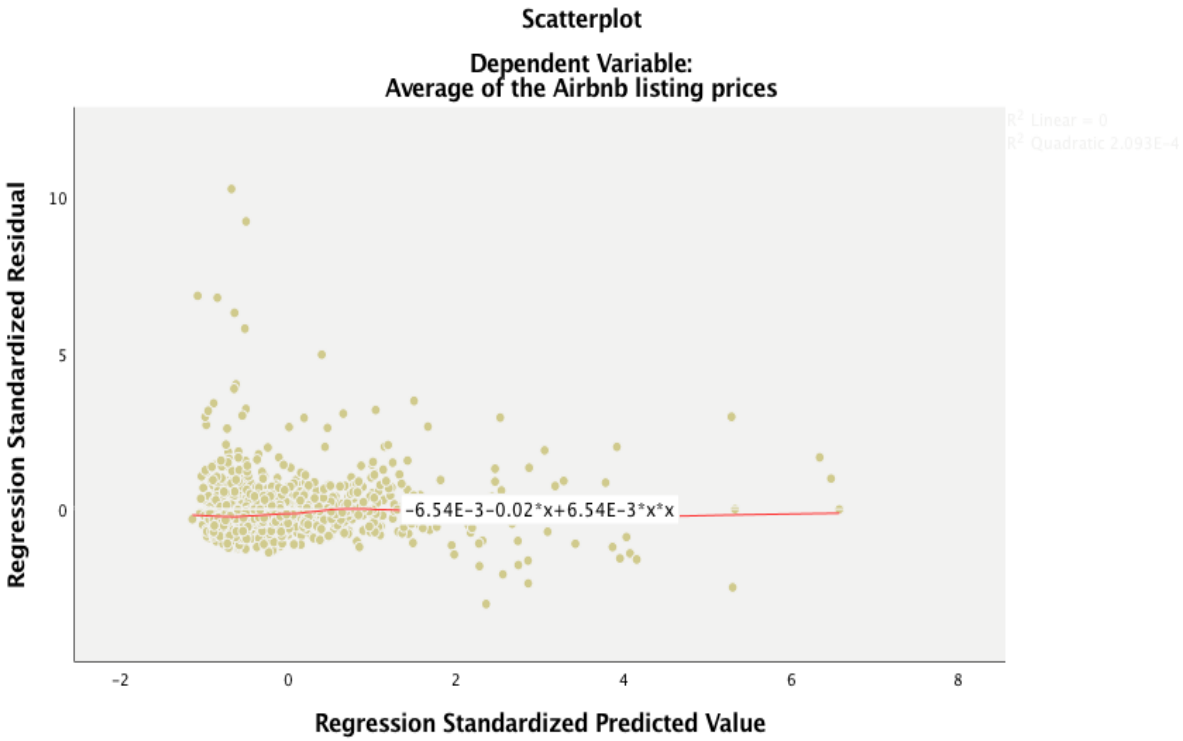
It's difficult to tell the relationship simply from this plot. Let's try fitting a non-linear best fit line (also known as Loess Curve) to see if we can detect any nonlinearity. You could double click on the scatterplot, then in the pop-up Output window choose:

`The icon "Add Fit Line at Total" or`

`right click of the dots -> Add Fit Line at Total`



Your scatterplot will now have a Loess curve fitted through it, which appears that the relationship of standardized predicted to residuals is roughly linear around zero, since the residuals seem to be randomly scattered around zero. This does not change the regression analysis.



1. Tests on Normality of Residuals

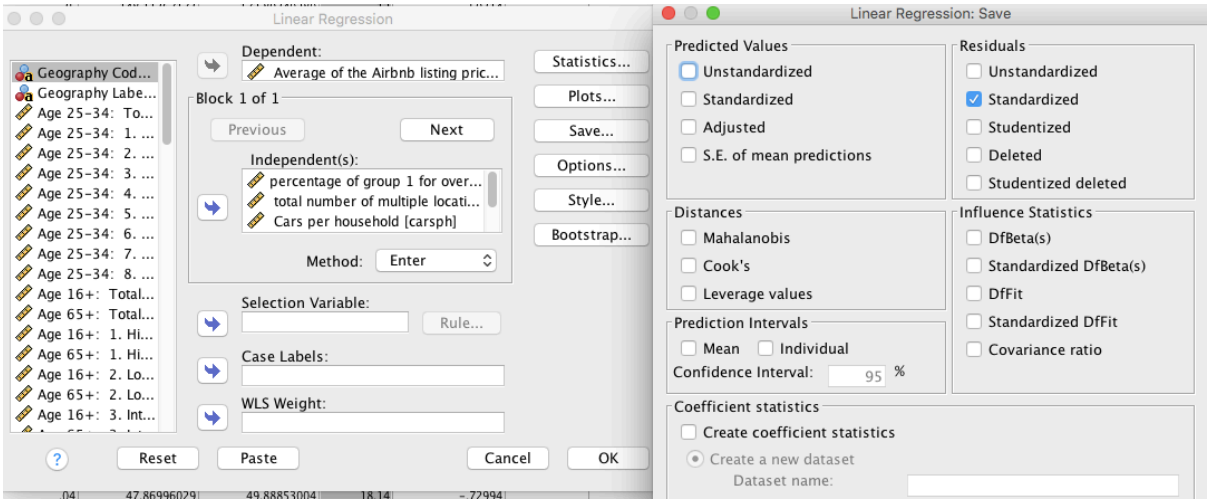
In linear regression, the outcome assumption is that the *residuals* are normally distributed, with the tested p-values for the t-tests being valid, in the format that errors distributed normally. The

normality test for residuals is necessary for the *b*-coefficient tests to be valid, which requires the errors be identically and independently distributed.

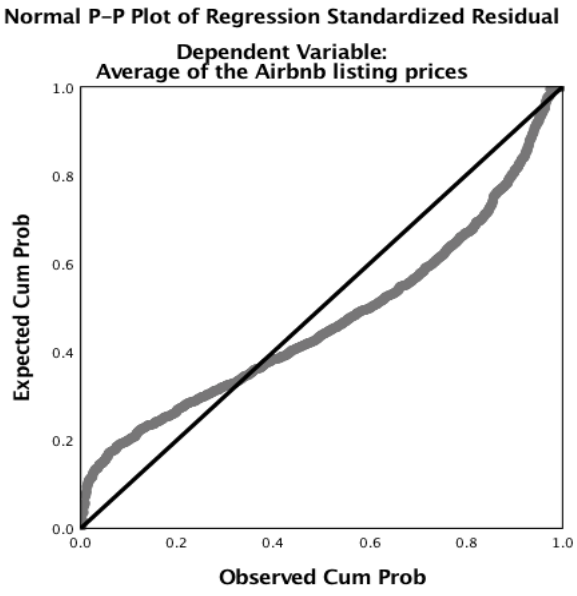
Let's go back and predict Airbnb listing price from Multiple Linear Regression model in page 4 of this practical (*Note*: normality of residuals test is model dependent, so it may change if more predictors added). On Page 5, we've already produced the standardized residuals plots by checking "Histogram" and "Normal Probability Plot" box. Hence you've got a preliminary idea of the normality distribution of model residuals. However, we still want to create the more commonly used Q-Q plot, which requires a further step to save the standardized residuals as a new variable (it will be automatically saved into the dataset as "ZRE_1").

`Analyse→ Regression → Linear`

click on **Save** and check Standardized under Residuals like below:



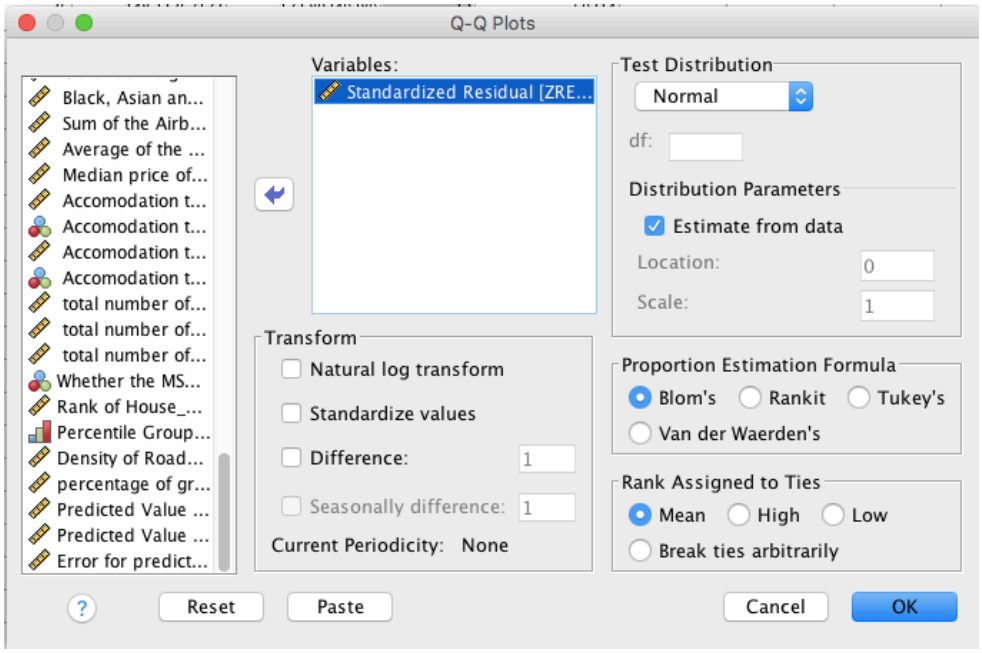
The P-P plot received (also in Page 5 in this practical) compares the observed cumulative distribution function (CDF) of the standardized residual to the expected CDF of the normal distribution.

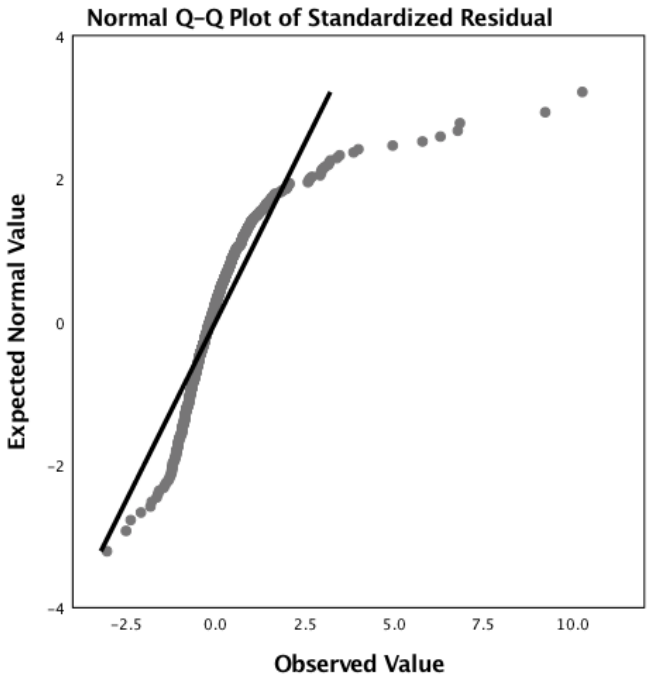


More commonly, Q-Q plots are better which compares the observed quantile with the theoretical quantile of a normal distribution; it is also more sensitive to tail distributions. To get the Q-Q plot for residuals, you need to

`Analyze – Descriptive Statistics – Q-Q Plots`

Choose the newly created variable ZRE_1 to the Variables box and click OK.





The resulting Q-Q plot indicates that, if it is a normal distribution then the points are expected to cluster around the horizontal line.

2. Check the Model Specification

The estimate of regression coefficients could be substantially affected by model specification. Commonly, a properly specified multiple linear regression model is expected to include all relevant variables, and exclude irrelevant variables; hence if relevant variables are omitted from the model or irrelevant variables are included in the model, the error term in the model may change due to wrongly attributed common variance shared among them.

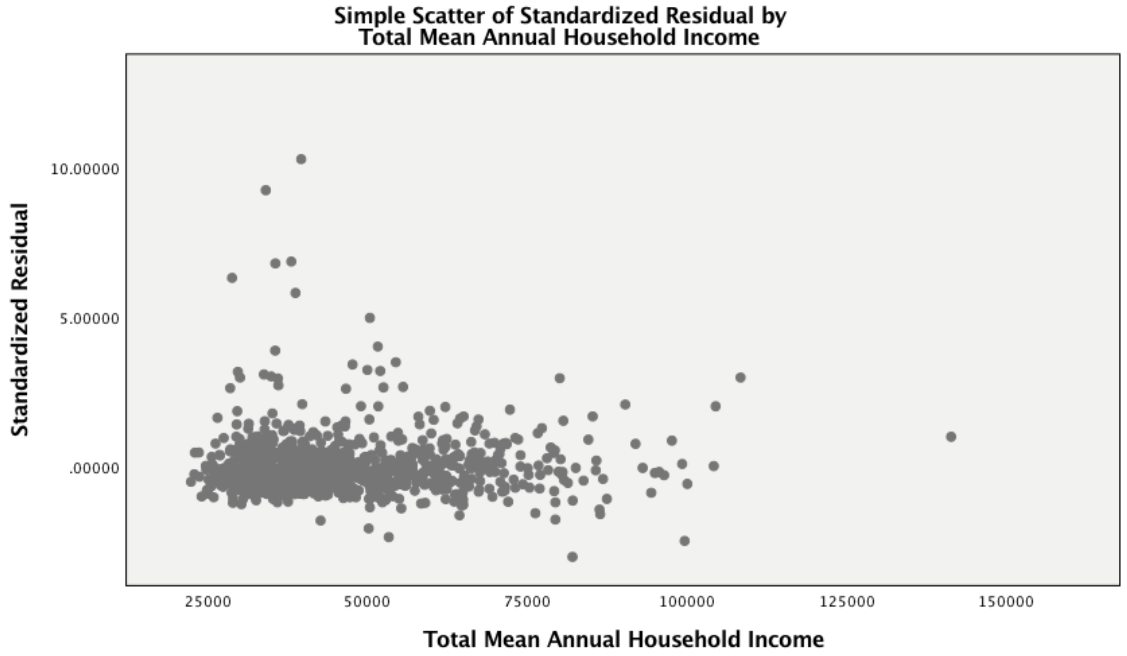
For this model, we could tell from the regression result that "Mean Annual Household Income" doesn't present significant relationship with Average Airbnb Price; before we delete the variable from the model, we would like to check the model specification.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	48.876	3.373		14.490	.000	42.257	55.495
	percentage of group 1 for over 16	101.705	31.510	.169	3.228	.001	39.870	163.541
	total number of multiple location hosts listed airbnbs	.178	.020	.228	8.754	.000	.138	.218
	Cars per household	-20.104	3.333	-.166	-6.031	.000	-26.645	-13.562
	Total Mean Annual Household Income	3.757E-5	.000	.014	.218	.828	.000	.000
	Average House price in Dec 2017	4.564E-5	.000	.437	11.545	.000	.000	.000

a. Dependent Variable:
Average of the Airbnb listing prices

As we've saved a new variable "ZRE_1", we will produce a scatterplot with "ZRE_1" (along the y-axis) with the "Mean Annual Household Income" on the x-axis. If in fact **Income** had no relationship with our model, it would be independent of the residuals.



From the graph, we can see that "Total Mean Annual Household Income" didn't show significant relationship with the residuals from our model, indicating independency from the residuals, hence the variable could be omitted from the model.

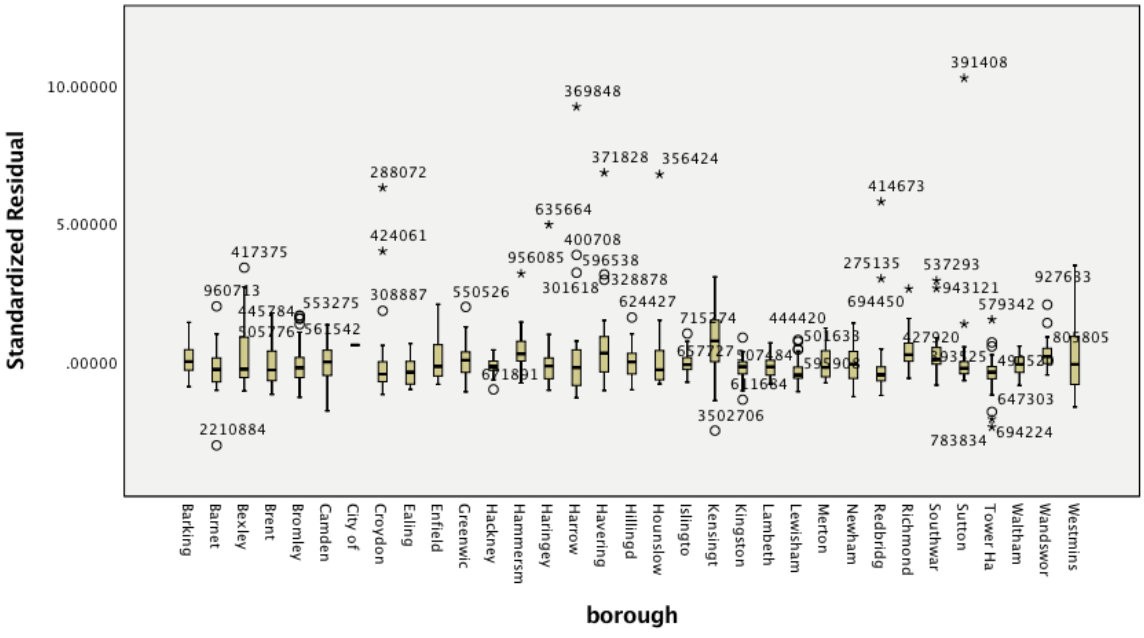
3. Issues of Independence

To check the independence of variables, the hypothesis assumption H_0 is that the errors associated with one observation are not correlated with the errors of any other observation; on the contrary, H_1 assumption is that their errors are not independent from each other. This model might violate the H_0 assumption whilst abide by the H_1 assumption, because considering the First Law of Geography, nearer places tend to have stronger influences on target area; when we collecting the data at MSOA level, it seems that MSOAs within the same Borough tend to be more like one another than those MSOAs from Boroughs far away from it, which working towards the errors.

To test the independence, we are going to create boxplots of the variable on standardized residuals (ZRE_1) clustered by Borough, to see if there is a pattern. Most notably, if the mean standardized residual is around zero for all boroughs, then we could conclude at the H_1 assumption.

`Graphs – Legacy Dialogs – Boxplot`

In the pop-up dialogue, click on ***Simple*** and ***Summaries for groups of cases***, then click on **Define**. Then choose: Summaries for Groups of Cases select Variable (ZRE_1), Category Axis (borough) and Label Cases by (Average House Price) respectively, then **OK**.

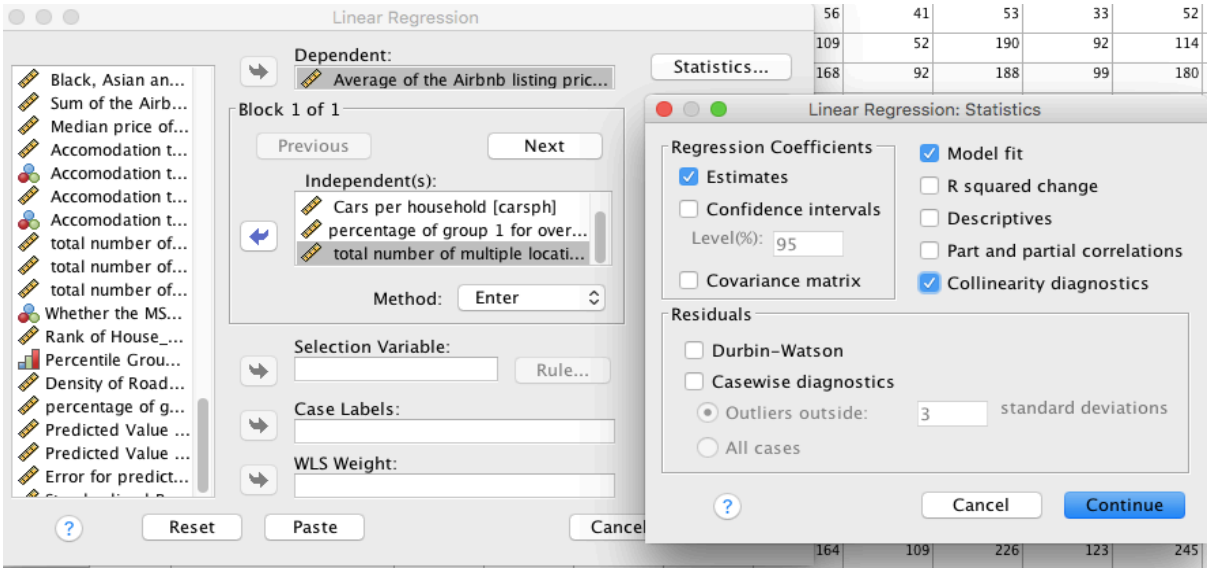


From the boxplot in output, boroughs tend to have similar mean residuals centered at zero, which suggests that the errors are not independent from each other.

4. Tests on Multicollinearity

Concerning about the above output on independency test, you might consider to further conduct multicollinearity test on the predictors. It refers to the possibility that predictors might be highly related to each other and all predictive of your outcome, which would cause problems in estimating the regression coefficients. For example, as the degree of multicollinearity increases, the coefficient estimates become unstable, the corresponding standard errors for coefficient estimates could be inflated dramatically.

Let's run the multiple linear regression again but adjust the model by ***omitting the variable on mean annual household income***, with Mean Airbnb listing price as dependent variable, percentage of Group 1 in all residents over 16, average house prices, multiple location Airbnb listings, average number of cars in each household as the input variables. By clicking on ***Statistics***, you will be able to check **Collinearity diagnostics** as below.



The Collinearity Tolerance specification is an indication of the percent of variance in the predictor that cannot be accounted for by the other predictors. This means that very small values indicate that a predictor is redundant, which means that values less than 0.10 are worrisome. The VIF, which stands for **variance inflation factor**, is (1/tolerance) and as a rule of thumb, a variable whose VIF values is greater than 10 are problematic.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.763[a]	.583	.581	26.2395249

a. Predictors: (Constant), total number of multiple location hosts listed airbnbs, percentage of group 1 for over 16, Cars per household, Average House price in Dec 2017

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	932117.749	4	233029.437	338.453	.000[b]
	Residual	667168.773	969	688.513		
	Total	1599286.52	973			

a. Dependent Variable: Average of the Airbnb listing prices

b. Predictors: (Constant), total number of multiple location hosts listed airbnbs, percentage of group 1 for over 16, Cars per household, Average House price in Dec 2017

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	49.230	2.953		16.669	.000		
	Average House price in Dec 2017	4.616E-5	.000	.442	14.642	.000	.472	2.121
	Cars per household	-19.716	2.818	-.162	-6.998	.000	.799	1.251
	percentage of group 1 for over 16	107.502	16.851	.178	6.380	.000	.550	1.818
	total number of multiple location hosts listed airbnbs	.178	.020	.227	8.768	.000	.640	1.562

a. Dependent Variable: Average of the Airbnb listing prices

In this example, with the multicollinearity eliminated, the coefficient for most of the predictors, which had been non-significant, is now significant.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions				
				(Constant)	Average House price in Dec 2017	Cars per household	percentage of group 1 for over 16	total number of multiple location hosts listed airbnbs
1	1	3.830	1.000	.01	.01	.01	.01	.01
	2	.846	2.128	.00	.00	.02	.00	.49
	3	.195	4.432	.05	.23	.13	.12	.36
	4	.078	7.010	.01	.74	.01	.87	.03
	5	.051	8.634	.93	.02	.83	.01	.11

a. Dependent Variable: Average of the Airbnb listing prices

Summary So Far

We assessed the assumptions of regression using SPSS, and the consequent diagnoses. As we have seen, it is not sufficient to simply run a regression analysis; upon various diagnosis of regression results, if your data does not meet the assumptions of linear regression, your coefficient estimates could be misleading, and corresponding interpretation could be in doubt; this may call our attention on possible spatial components and relevant interpretation in next week.