Exercises

Statistical Tests & Linear Regression

By Anna Meudec

Statistical Tests and Linear Regression

The goal of these exercises is to practice doing statistical tests and developing regression models in Jamovi. The dataset used for these exercises can be found in the "Files Employed" section at the end of these exercises.

Note: Prior to performing the statistical tests and linear regression below, the following outlier has been removed.

House Code: 2001

Price: €9,000,000

Part A.

Generating univariate descriptive statistics for variables - house price, house size, number of rooms and number of bathrooms.

House Price

Descriptives

	Price
N	249
Mean	598305
Median	605000
Standard deviation	151356
Minimum	228000
Maximum	1051000

House Size

Descriptives

	House Size (Sq. M)
N	249
Mean	132
Median	135
Standard deviation	21.7
Minimum	66
Maximum	173

Number of Rooms

Descriptives

	No. of Rooms
N	249
Missing	0
Mean	6.59
Median	7
Standard deviation	1.15
Minimum	3
Maximum	9

Number of Bathrooms

Descriptives

	No. of Bathrooms
N	249
Mean	1.92
Median	2
Standard deviation	0.460
Minimum	1
Maximum	3

Examining the correlations between all 4 variables.

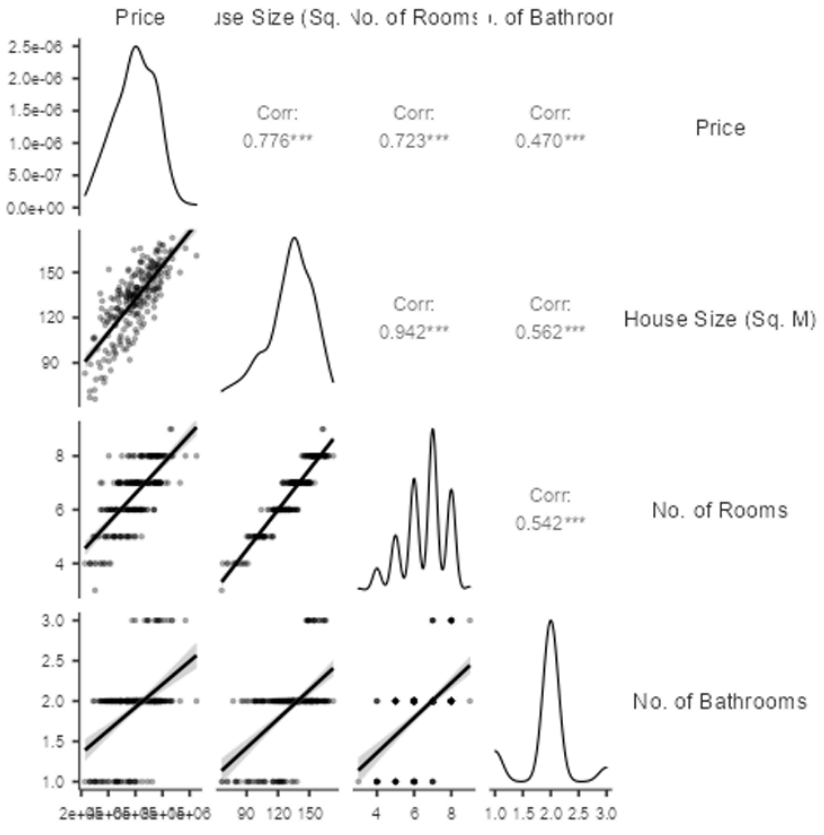
Correlation Matrix

Correlation Matrix

		Price	House Size (Sq. M)	No. of Rooms	No. of Bathrooms
Price	Pearson's r	—			
	p-value	—			
	95% CI Upper	—			
	95% CI Lower	—			
	N	—			
House Size (Sq. M)	Pearson's r	0.776***	—		
	p-value	< .001	—		
	95% CI Upper	0.821	—		
	95% CI Lower	0.721	—		
	N	249	—		
No. of Rooms	Pearson's r	0.723***	0.942***	—	
	p-value	< .001	< .001	—	
	95% CI Upper	0.778	0.955	—	
	95% CI Lower	0.658	0.926	—	
	N	249	249	—	
No. of Bathrooms	Pearson's r	0.470***	0.562***	0.542***	—
	p-value	< .001	< .001	< .001	—
	95% CI Upper	0.562	0.641	0.624	—
	95% CI Lower	0.367	0.470	0.447	—
	N	249	249	249	—

Note. * p < .05, ** p < .01, *** p < .001

Plot



Part B.

Using the House Price Data to perform the following statistical tests:

(i) Is there a significant difference between the price of a house in the East and West?

Price by Location (Independent Samples T-Test)

Independent Samples T-Test

Independent Samples T-Test

		Statistic	df	p
Price	Student's t	-3.62	247	< .001

Assumptions

Normality Test (Shapiro-Wilk)

	W	p
Price	0.992	0.228

Note. A low p-value suggests a violation of the assumption of normality

Homogeneity of Variances Test (Levene's)

	F	df	df2	p
Price	0.0600	1	247	0.807

Note. A low p-value suggests a violation of the assumption of equal variances

[3]

Group Descriptives

	Group	N	Mean	Median	SD	SE
Price	West	97	555845	572000	146354	14860
	East	152	625401	627000	148694	12061

Null Hypothesis: There is no difference between the price of a house in the East and West. The reported p-value of this test is <.001. Therefore, we can reject the null and accept the

alternative that there is a significant difference between the price of a house in the East and West.

To be more specific, we can see from the mean values in the table above that houses in the East have an average reported price of €625,401, whilst, in the West, houses have an average reported price of €555,845.

Independent Sample T-Test Assumptions:

When running the independent sample t-test, a number of assumptions were made.

1. I assumed that the data for both price and location were independent observations selected from a representative random sample of the population.
2. I also assumed that the test variable (price) is a ratio variable that follows a normal distribution.
3. Lastly, I assumed that the variances of both price and location were both equal.

In the examination of assumption 2, I performed the Shapiro-Wilks test for normality (p=.001). On completion of this test, the value deduced was p = 0.228. This suggests that the null may be correct. However, we are using large sampling distributions (> 50), so the assumption that the sample means are normal will hold.

The next test I performed was to check for the equality of variances using Levene's homogeneity of variance (p=.131 > .05). The result of the test was p = 0.807, which suggests it's not significant. Therefore, there is no significant difference in the variability of the two distributions. So, the assumption of equal variances still holds.

(ii) Is there a significant difference between the price of a terraced, semi-detached and detached house?

Price by House Type (One-Way ANOVA)

One-Way ANOVA

One-Way ANOVA (Welch's)

	F	df1	df2	p
Price	120	2	92.4	< .001

Group Descriptives

	House Type	N	Mean	SD	SE
Price	Detached	128	686883	115942	10248
	Semi Detached	89	544640	111273	11795
	Terrace	32	393250	94766	16752

Null Hypothesis: There is no difference between the prices of terraced, semi-detached, and detached houses.

The reported p-value of this test is <.001. Therefore, we can reject the null and accept the alternative that there is a significant difference between the prices of the three types of houses.

(iii) Is there a significant difference between the types of houses in the East and the West?

Type of House by Location (Chi-Square Test)

Contingency Tables

Contingency Tables

House Type		Location		
		West	East	Total
Detached	Observed	54	74	128
	% within column	55.7 %	48.7 %	51.4 %
Semi Detached	Observed	34	55	89
	% within column	35.1 %	36.2 %	35.7 %
Terrace	Observed	9	23	32
	% within column	9.3 %	15.1 %	12.9 %
Total	Observed	97	152	249
	% within column	100.0 %	100.0 %	100.0 %

χ^2 Tests

	Value	df	p
χ^2	2.16	2	0.339
N	249		

Note: Chi-Square = 2.16, p = 0.339, therefore p > 0.05

Null Hypothesis: There is no difference between the types of houses in the East and the West.

The p-value produced from the test is >.05, so in this case, we cannot reject the null hypothesis and cannot accept the alternative that there is a significant difference between East and West.

However, just because we cannot prove that the null hypothesis is incorrect, it doesn't mean the null hypothesis remains true, we simply don't have enough evidence to prove otherwise.

(iv) Do consumers believe that House Size or House Price is more important when buying a house?

Paired Samples T-Test

Paired Samples T-Test

								95% Confidence Interval			
			statistic	df	p	Mean difference	SE difference	Lower	Upper		Effect Size
Price	House Size (Sq. M)	Student's t	62.4	248	< .001	598173	9591	579284	617063	Cohen's d	3.95

Normality Test (Shapiro-Wilk)

			W	p
Price	-	House Size (Sq. M)	0.991	0.123

Note. A low p-value suggests a violation of the assumption of normality

Descriptives

	N	Mean	Median	SD	SE
Price	249	598305	605000	151355.6	9591.77
House Size (Sq. M)	249	132	135	21.7	1.38

Null Hypothesis: There is <u>no difference</u> between the mean of both variables, house size and house price.

The reported p-value of the Paired Sample t-test is <.001. Therefore, we can reject the null hypothesis and accept the alternative, that there is a significant difference between the mean of the variables, house size and house price. The mean difference indicates that there is quite a big difference in both variables' means', and similarly, the effect size (Cohen's D) indicates a very large effect size, with standard deviations between the variables. Furthermore, through obtaining confidence intervals for the differences using descriptive plots on Jamovi, it can be seen that the mean and confidence interval for both variables are extremely far away from each other, with absolutely no overlap.

Paired Sample T-Test Assumptions:

When running the paired sample t-test, the following assumptions were made.

1. Both variables are continuous (ratio/interval).
2. The data for both variables are independent observations selected from a representative random sample of the population.
3. The distribution of differences is normally distributed.

In order to test assumption 3, I conducted the Shapiro-Wilks test for normality (p=.001). The test resulted in p = 0.123, suggesting that the null hypothesis may, in fact, be correct. A normal distribution is necessary to perform a valid t-test. However, in this example, we are working with a large sampling distribution (>50), so the assumption that the sample means are normal will still hold in this case.

Part C. Linear Regression

Developing a regression model in Jamovi with house price as the dependent variable.

Regression Model 1

Model Fit Measures

Model	R	R^2	Adjusted R^2
1	0.804	0.647	0.636

Model Coefficients - Price

Predictor	Estimate	SE	t	p
Intercept [a]	147901	168194	0.879	0.380
House Type:				
Semi Detached – Detached	−84477	38286	−2.206	0.028
Terrace – Detached	−114350	40677	−2.811	0.005
Location:				
East – West	48314	14666	3.294	0.001
House Size (Sq. M)	4206	1247	3.372	< .001
No. of Rooms	−6164	15183	−0.406	0.685
No. of Bathrooms	16355	15517	1.054	0.293
Land Area (Hectares)	−114128	356370	−0.320	0.749

[a] Represents reference level

In **Regression Model 1** on the right, I have used all independent variables that I believe could have an influence on house prices. The model does, however, exclude our satisfaction data due to it being ordinal.

In the model, it is clear that many of the independent variables are non-significant and affect model fit. In order to improve the regression model, I will run tests and assumptions to see which variables are causing errors in the model.

Multicollinearity

In order to examine the correlation between independent variables in the model, I created a correlation matrix and also performed the assumption check for collinearity.

In the correlation matrix, it was found that the following variables had strong positive correlations.

- Land Area & House Size (0.867)
- Land Area & No. of Rooms (0.813)
- No. of Rooms & House Size (0.942)

Furthermore, on testing collinearity, as expected, low tolerances were seen for No. of Rooms, House Size and Land Area. So, by removing the No. of Rooms data and also Land Area data, the model will begin to fit better, and estimates of the b coefficients will be more accurate (view regression model 2). I will not remove house size, however, as it is a highly important predictor of our dependent variable (Price).

Regression Model 2

Model Fit Measures

Model	R	R^2	Adjusted R^2
1	0.804	0.646	0.639

Model Coefficients - Price

Predictor	Estimate	SE	t	p
Intercept [a]	97923	64529	1.52	0.130
House Type:				
Semi Detached – Detached	−73417	15197	−4.83	< .001
Terrace – Detached	−105530	30462	−3.46	< .001
Location:				
East – West	46340	12897	3.59	< .001
House Size (Sq. M)	3648	510	7.15	< .001
No. of Bathrooms	15820	15424	1.03	0.306

[a] Represents reference level

Residual Analysis

I used residual plots on Jamovi to be able to look for patterns in the residuals. Through performing the residual analysis, I noticed that the residuals for the No. of Bathrooms followed a pattern, indicating that an assumption has been violated. Therefore, I removed this independent variable from the model.

Final Regression Model

Model Fit Measures

Model	R	R^2	Adjusted R^2
1	0.803	0.645	0.639

Model Coefficients - Price

Predictor	Estimate	SE	t	p
Intercept [a]	97526	64535	1.51	0.132
House Type:				
Semi Detached – Detached	-73522	15198	-4.84	< .001
Terrace – Detached	-101510	30212	-3.36	< .001
Location:				
East – West	46543	12897	3.61	< .001
House Size (Sq. M)	3876	459	8.45	< .001

[a] Represents reference level

Interpreting b-values

X1: House Size (Sq. M): b1 = €3,876

An increase of 1 Sq. M corresponds to an average increase in price of €3,876.

X2: Location (East-West): b2 = €46,543

Houses in the East are, on average, €46,543 more than Houses in the West.

X3: House Type: Semi-Detached – Detached: b3 = -€73,522

A Semi-Detached house is, on average, €73,522 less than a detached house.

X4: House Type: Terrace – Detached: b4 = -€101,510

A Terrace house is, on average, €101,510 less than a detached house.

b-values Significance

House Size and Location have statistically significant positive b coefficients, and House Types have statistically significant negative b coefficients.

Model Fit R-Square = .645

64.5% of the variations in House Prices are explained by the independent variables in the model, indicating that the data fits the model very well.

References

[1] The jamovi project (2021). jamovi. (Version 2.0) [Computer Software]. Retrieved from https://www.jamovi.org.

[2] R Core Team (2021). R: A Language and environment for statistical computing. (Version 4.0) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2021-04-01).

[3] Fox, J., & Weisberg, S. (2020). car: Companion to Applied Regression. [R package]. Retrieved from https://cran.r-project.org/package=car.

Files Employed

CSV House Price Dataset: **house_price_dataset.csv (Google Drive)**

OMV Jamovi File: **statistical_tests_&_linear_regression.omv (Google Drive)**