TARTU ÜLIKOOL

Multiple regression as an evaluation tool for the study of relationships between energy consumption, air-tightness and their influencing factors

Presented by **Cagatay Ipbüker**, M.Sc.

Authors:

C. Ipbüker & A.H. Tkaczyk

University of Tartu, Institute of Physics, Tartu, Estonia

M. Valge & K. Kalbe & T. Mauring

University of Tartu, Institute of Technology, Tartu, Estonia

Scope of study

- We examine the relationship between energy consumption, air-tightness and their influencing factors including volume, treated floor area, building age, compactness, window to wall ratio and frame length factor of 18 dwellings (all built post-2000).

Aim of study

- We employ one-way ANOVA analysis to compare the means of air-tightness and energy consumption in relation to other individual factors as well as partial least squares and principal component analysis to investigate dominating variables, trends, groups and similarities and also to find the relationships between two sets of multivariate data. We then utilize multiple regression to explore the relationship between all parameters
- The findings from the developed predictive model for estimating the energy consumption of dwellings with the multivariate regression technique should contribute in predicting the influence of design parameters on achieving air-tightness and energy efficiency on buildings more easily and routinely.

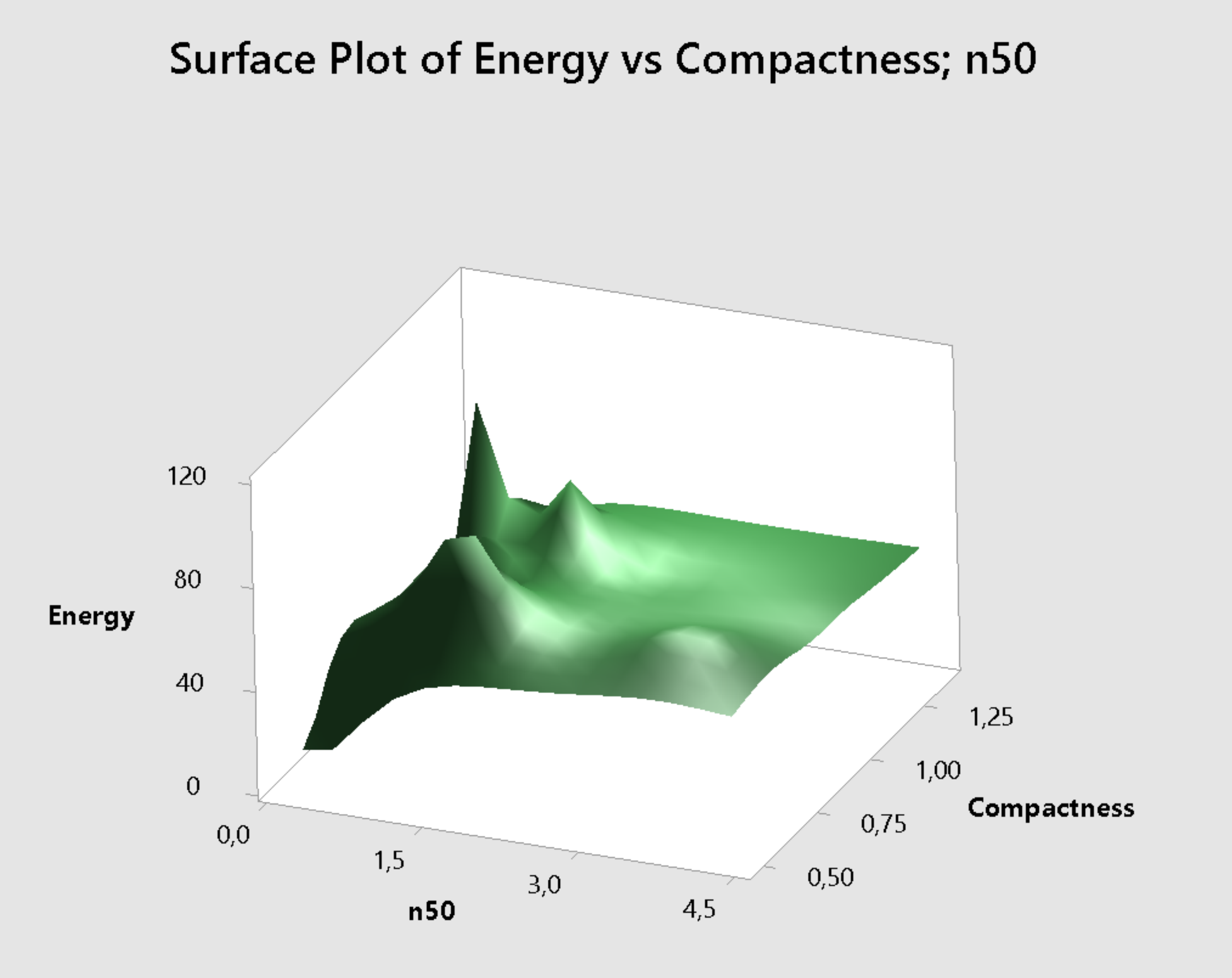
Data

V	TFA	Heat Demand	n50	Age	Compactness	WWR	FLF
(m3)	(m2)	(kWh/(m2*a)	(1/h)	(year)	(m2/m3)	(m2/m2)	(m/m3)
400,00	151,70	14,50	0,25	0,83	1,14	0,13	0,19
1586,40	337,00	14,61	0,36	2,00	0,49	0,17	0,10
400,00	151,70	14,90	0,33	0,40	1,14	0,13	0,19
430,00	152,70	15,00	0,42	1,00	1,20	0,18	0,32
708,00	263,00	15,00	0,30	0,40	1,00	0,20	0,32
425,00	151,70	21,00	0,42	0,40	1,10	0,11	0,25
410,00	149,70	22,00	0,34	3,00	0,97	0,18	0,23
340,00	135,00	24,00	0,39	1,00	1,36	0,32	0,66
450,00	150,00	30,00	0,80	0,00	0,96	0,22	0,28
270,00	130,00	40,00	0,96	0,00	1,13	0,15	0,23
710,00	210,00	45,00	2,50	4,00	1,08	0,27	0,47
622,00	218,00	53,00	4,40	10,00	0,97	0,17	0,37
395,00	152,00	66,00	3,60	5,00	0,68	0,17	0,20
515,00	141,00	67,50	0,95	4,00	0,82	0,46	0,49
378,00	144,00	68,00	0,79	4,00	1,08	0,18	0,32
708,00	248,00	75,00	1,40	4,00	1,16	0,20	0,26
950,00	252,00	90,00	1,30	7,00	0,74	0,27	0,24
614,20	296,00	117,00	0,60	3,00	1,18	0,17	0,36

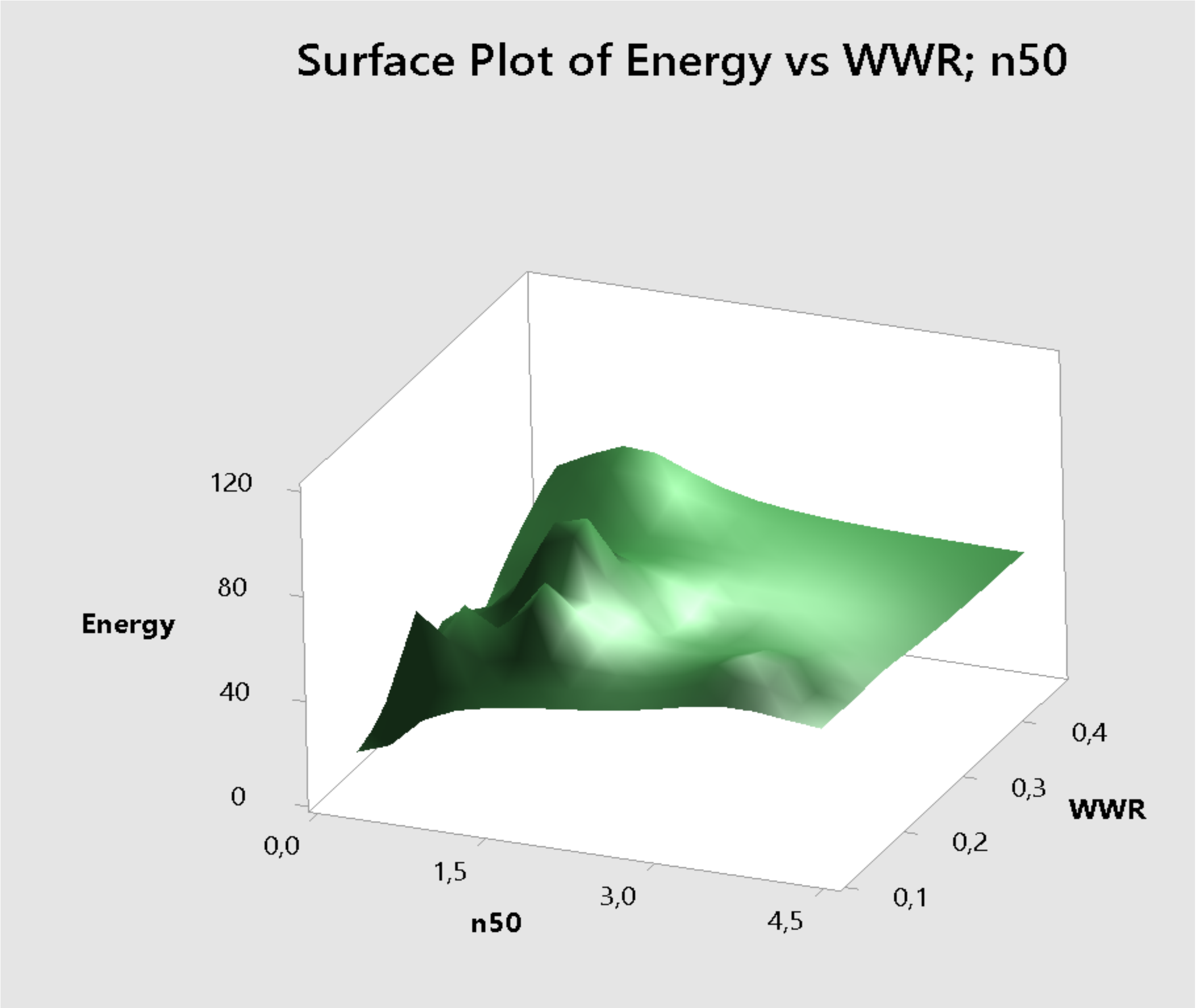
Example (Sample Building & Compactness)

- Almost square shape, compactness equals to 1,14 m^2/m^3.

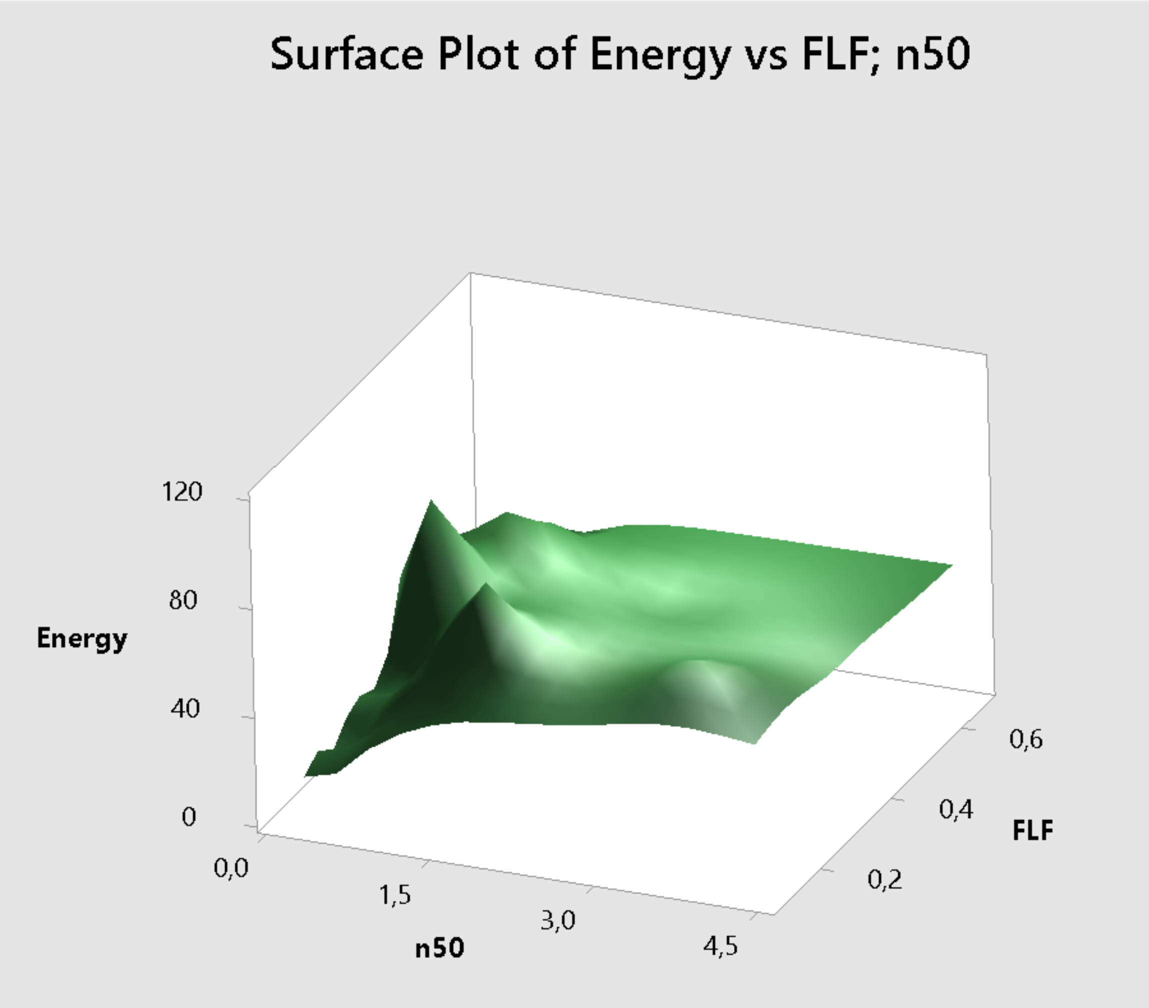
Cross examination of variables



Cross examination of variables



Cross examination of variables



The multiple linear model

- The study of P-values for each independent variables for the built models reveals that each independent variable is important to keep in the model.
- However, the analysis of the residuals (difference of values between the measured energy and the predictive (based on the model)) shows that the houses with high energy consumption have higher residuals.
- That's why the model built on data obtained from „more“ energy efficient houses (<60 (kWh/(m2*a)) is proven to be a better fit.

Models

(For the whole data, all independent parameters included)

- Energy= -25,5-0,1324*V+0,599*TFA-2,74*n50+6,49*Age-0,7*Compactness+239*WWR-106*FLF

R^2_adjusted of 47,54%

R^2 of 69,14%

(For Energy<60 (kWh/(m2*a)), 4 independent parameters)

- Energy = -18,5+11,2*n50+26,4*Compactness+134,3*WWR-63,2*FLF

R^2_adjusted of 75,07%

R^2 of 84,14%

(For Energy<60 (kWh/(m2*a)), all independent parameters)

- Energy=69,0-0,0077*V-0,0837*TFA+14,79*n50-2,89*Age-40,4*Compactness+16,6*WWR+23,0*FLF

R^2_adjusted of 83%

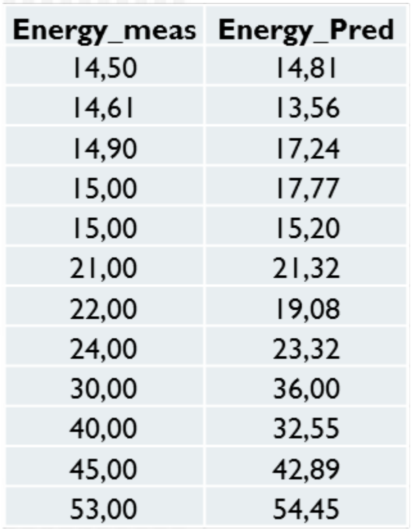
R^2 of 93,85%

Best fit

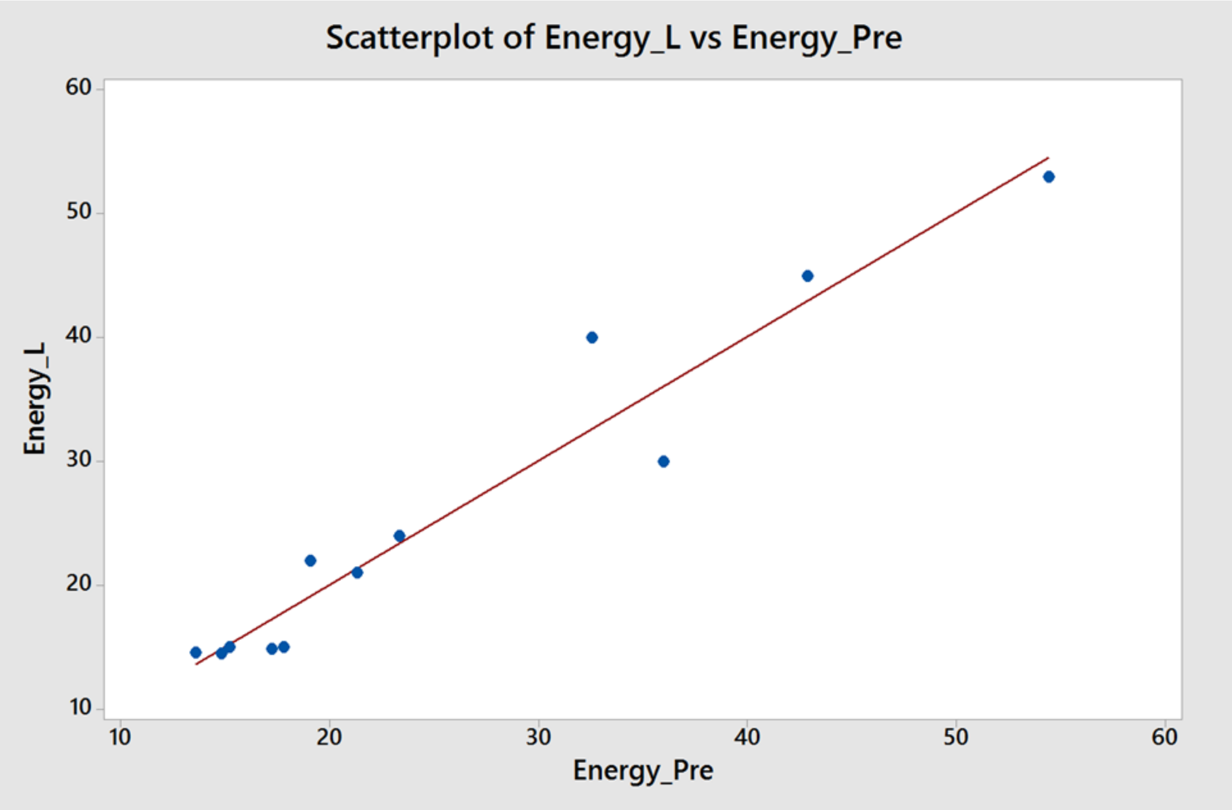
- Energy=69,0-0,0077*V-0,0837*TFA +14,79*n50-2,89*Age-40,4*Compactness+16,6*WWR+23,0*FLF

R^2_adjusted of 83%

R^2 of 93,85%



Energy_meas	Energy_Pred
14,50	14,81
14,61	13,56
14,90	17,24
15,00	17,77
15,00	15,20
21,00	21,32
22,00	19,08
24,00	23,32
30,00	36,00
40,00	32,55
45,00	42,89
53,00	54,45



Scatterplot of Energy_L vs Energy_Pre

Some explanation

1) What is R^2 and R^2-adjusted?

R-squared is known as the coefficient of determination (or multiple determination in our case)

R^2 is basically the percentage of response variable variation in the model that is explained by its relationship with independent variables.

Very simply: higher the number of variables, higher the R^2

So, we have R^2-adjusted because it allows of to compare regressions with different numbers of variables.

2) What is the p-value?

Determines the appropriateness of rejecting the null hypothesis in a significance test. P-values range from 0 to 1. In our work, p-value is mostly used in the „lack-of-fit" test, which assess the fit of the model and check for replicates.

Variables and P-value

y = ß0 + ß1x1 + ß2x2 + ß3x3 + … + ß8x8 + …+ ß12x12
+ ß13x13 + ß14x14 + …+ ß32x32 + e

We can get rid of the excess-variables with 3 basic approaches:

1)Forward selection

2)Backward selection

3)Mixture of the two

They all come to the same conclusion: Drop the variable with the least „significant“ variation (meaning a high P-Value), but after so many tests, the P-Values may become compromised.

Problems?

So far, I have received 2 important questions:

1) Why multiple linear regression?
2) Too small sample size?

Problems?

1) Why multiple linear regression?

We are often interested if a dependent variable is related to more than one independent variable:

$$y = ß0 + ß1x1 + e$$

$$y = ß0 + ß2x2 + e$$

$$y = ß0 + ß3x3 + e$$

This is a possible solution, but when we test in such a „stepwise“ way, we are putting the variables in privileged positions.

Multiple linear regression allows us simultaneous testing and modeling of independent variables.

Correlation does not mean causation

2) Too small sample size?

We can even generate correct results with as low as 3 cases, because we are not dealing with observational data (social sciences), but with measured physical variables, so we already know that a physical relationship exists.

Statistically, our sample size is small, because our statistical power is low. In such a case, it’s recommended not to use ANOVA / ANCOVA / MANOVA /MANCOVA as well as multiple regression analysis, as the small sample will generate biased results. Fischer’s exact test or Chi-Squared test are better to use.

Adequate sample size should be preferable to large samples as „overfitting“ will manifest itself which will lead the model to reveal biased results.

Tänan tähelepanu eest!

cagatay@ut.ee