Statistics Kingdom

Multiple Linear Regression Calculator

Multiple regression calculator with unlimited predictors.

Tutorial Information Simple linear regression Regression sample size

Iterations:

Automatically

Significance level (α):

0.05

Effect:

Medium

Effect type:

f

Effect size:

0.39

Digits:

6

☐ **Constant is zero** (Force zero Y-intercept, b_0=0)

☐ **Power regression** - Ln transformation (natural log) over all the variables: $Y=\exp(b_0)\cdot X_1^{b_1}\cdot\cdot X_p^{b_p}$.

◉ Enter raw data directly
○ Enter raw data from excel

*Include	✔	✔	✔	✘	✘	✘	
Transform							
Groups	X1	X2	X3	X4	X5	X6	Y
Data	0 0 0 0 0 0 2 2 2 3 3 1 4 4 4	0 0 0 0 0 0 0 0 0 0 0 1 0 0 0	0 0 0 0 0 0 0 0 0 0 0 2 0 1 1	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	10 2 9 8 10 4 9 19 15 12 15 8 12 13 17	12 11 10 10 3 3 10 3 3 10 3 10 3 10 3	0 0 0 0 0 0 2 2 2 3 3 4 4 5 5
P-value:	1.11022e-16	-2.22045e-16	2.22045e-16				
Average:	177.59	7.482	27.052				212.12
n:	500	500	500				500
S:	323.172528	21.18091	42.509286				380.233118
Skewness:	8.959056	11.767564	12.165423				9.712906
Normality:**	0	0	0				0
Outliers:	400, 407, 411, 412,	17, 17, 17, 17, 18,	68, 70, 70, 72, 74,				470, 472, 490, 494,

Pred Y	Residual
-0.00601575	0.00601575
-0.00601575	0.00601575
-0.00601575	0.00601575
-0.00601575	0.00601575
-0.00601575	0.00601575
-0.00601575	0.00601575
1.993984	0.0060165
1.993984	0.0060165
1.993984	0.0060165
2.993983	0.00601687
2.993983	0.00601687
3.994084	0.00591615
3.993983	0.00601724
4.994093	0.00590691
4.994093	0.00590691

212.12	0
500	500
380.233108	0.0894023
9.713052	-22.330457
0	0
469.9967603058015,	-1.99419822014762,

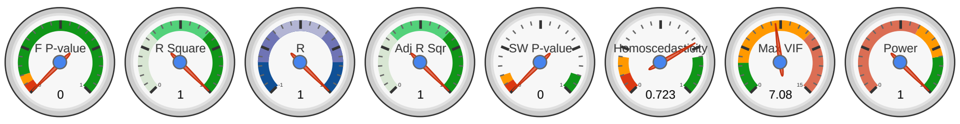
Calculate Insert column Delete column Clear Load last run Load example

How to do with R?

$$\hat{Y} = -0.00601575 + 1\ X_1 + 0.999879\ X_2 + 1.00011\ X_3$$

Reporting results in APA style

Results of the multiple linear regression indicated that there was a very strong collective significant effect between the X1, X2, X3, X4, X5, X6, and Y, ($F(3, 496) = 2990628758$, $p < .001$, $R^2 = 1$, $R^2_{adj} = 1$).
The individual predictors were examined further and indicated that X1 (t = 33293.746, $p < .001$) and X2 (t = 2534.48, $p < .001$) and X3 (t = 3980.233, $p < .001$) were significant predictors in the model, and were non significant predictors in the model.



Correlation matrix (pearson)

	Y	X_1	X_2	X_3	X_4	X_5	X_6
Y	1	0.997706	0.864944	0.928725	0.127307	-0.0274206	0.0183621
X_1	0.997706	1	0.837901	0.90425	0.127751	-0.0245564	0.0246168
X_2	0.864944	0.837901	1	0.868311	0.0278424	0.000921755	0.00617353
X_3	0.928725	0.90425	0.868311	1	0.153542	-0.0589944	-0.0259232
X_4	0.127307	0.127751	0.0278424	0.153542	1	-0.107383	-0.12769
X_5	-0.0274206	-0.0245564	0.000921755	-0.0589944	-0.107383	1	-0.17639
X_6	0.0183621	0.0246168	0.00617353	-0.0259232	-0.12769	-0.17639	1

ANOVA table

Source	DF	Sum of Square	Mean Square	F Statistic	P-value
Regression (between $\hat{y}_i$ and $\bar{y}$)	3	72144030.81	24048010.27	2990628758	0
Residual (between y_i and $\hat{y}_i$)	496	3.988396	0.00804112		
Total (between y_i and $\bar{y}$)	499	72144034.8	144577.224		

Coefficient Table Iteration 1 (adjusted R-squared = 1)

	Coeff	SE	t-stat	lower $t_{0.025}(493)$	upper $t_{0.975}(493)$	Stand Coeff	p-value	VIF
b	-0.00432255	0.0174249	-0.248068	-0.0385588	0.0299137	0	0.804185	
X1	1.000001	0.0000303411	32958.60528	0.999942	1.000061	0.849934	0	5.945197
X2	0.999973	0.000405872	2463.76146	0.999175	1.00077	0.0557036	0	4.569854
X3	1.000043	0.00025909	3859.829367	0.999534	1.000552	0.111803	0	7.500711
X4	0.00390668	0.00495258	0.788817	-0.00582409	0.0136374	0.00000876074	0.430598	1.102716
X5	-0.000435573	0.000950984	-0.458023	-0.00230405	0.00143291	-0.0000050044	0.647138	1.067241
X6	-0.000679223	0.0012595	-0.539281	-0.00315387	0.00179542	-0.00000590908	0.589937	1.073355

Coefficient Table Iteration 2 (adjusted R-squared = 1)

	Coeff	SE	t-stat	lower $t_{0.025}(494)$	upper $t_{0.975}(494)$	Stand Coeff	p-value	VIF
b	-0.00913875	0.0138834	-0.65825	-0.0364166	0.018139	0	0.510684	
X1	1	0.000030247	33061.14301	0.999941	1.00006	0.849933	0	5.917832
X2	0.999959	0.000404407	2472.655143	0.999164	1.000754	0.0557028	0	4.544182
X3	1.000057	0.000256904	3892.734002	0.999553	1.000562	0.111804	1.11022e-16	7.386474
X4	0.0041475	0.00492065	0.842877	-0.00552048	0.0138155	0.00000930079	0.399705	1.090287
X6	-0.000562134	0.0012323	-0.456169	-0.00298332	0.00185905	-0.00000489043	0.648469	1.029136

Coefficient Table Iteration 3 (adjusted R-squared = 1)

	Coeff	SE	t-stat	lower $t_{0.025}(495)$	upper $t_{0.975}(495)$	Stand Coeff	p-value	VIF
b	-0.0136226	0.00979691	-1.390497	-0.0328712	0.00562608	0	0.165003	
X1	0.999999	0.0000300537	33273.74723	0.99994	1.000058	0.849932	0	5.851794
X2	0.999958	0.000404079	2474.659433	0.999164	1.000752	0.0557027	0	4.544086
X3	1.000068	0.000255627	3912.221622	0.999566	1.00057	0.111806	0	7.324945
X4	0.00441525	0.00488161	0.904467	-0.00517598	0.0140065	0.00000990122	0.366188	1.074774

Coefficient Table Iteration 4 (adjusted R-squared = 1)

	Coeff	SE	t-stat	lower $t_{0.025}(496)$	upper $t_{0.975}(496)$	Stand Coeff	p-value	VIF
b	-0.00601575	0.00502367	-1.197483	-0.015886	0.00385453	0	0.23169	
X1	1	0.0000300357	33293.74562	0.999941	1.000059	0.849932	1.11022e-16	5.84692
X2	0.999879	0.000394511	2534.479639	0.999104	1.000654	0.0556983	-2.22045e-16	4.33302
X3	1.00011	0.000251269	3980.23334	0.999617	1.000604	0.11181	2.22045e-16	7.079951

Multiple linear regression

The backward stepwise method is used to produce an initial screening of the predictors. For the final independent variables scope, you need to incorporate your expertise.

1. Y and X relationship

R square (R^2) equals **1.** It means that the predictors (X_i) explain 100% of the variance of Y.
Adjusted R square equals **1.**
The coefficient of multiple correlation (R) equals 1. It means that there is a very strong correlation between the predicted data (ŷ) and the observed data (y).

2. Goodness of fit

Overall regression: right-tailed, $F_{(3,496)}$ = **2990628758**, p-value = **0**. Since p-value < α (0.05), we reject the H_0.
The linear regression model, **$Y = b_0 + b_1X_1 + ... + b_pX_p + \varepsilon$**, provides a better fit than the model without the independent variables resulting in, **$Y = b_0 + \varepsilon$**.

The following independent variables are not significant as predictors for Y: **X_5 X_6 X_4**.
Therefore the calculator excluded these variables from the model.

If any excluded variable is highly suspected to be related to the dependent variable (Y), theoretically or due to previous research, it is recommended to include the variable in the model irrespective of the p-value, to do it, you should change the iterations to **manual**.

The Y-intercept (b): two-tailed, T = **-1.197483**, p-value = **0.23169**. Hence b is not significantly different from zero. It is still most likely recommended not to force b to be zero.

If you like the page, please share or like. Questions, comments and suggestions are appreciated. (statskingdom@gmail.com)

Validation

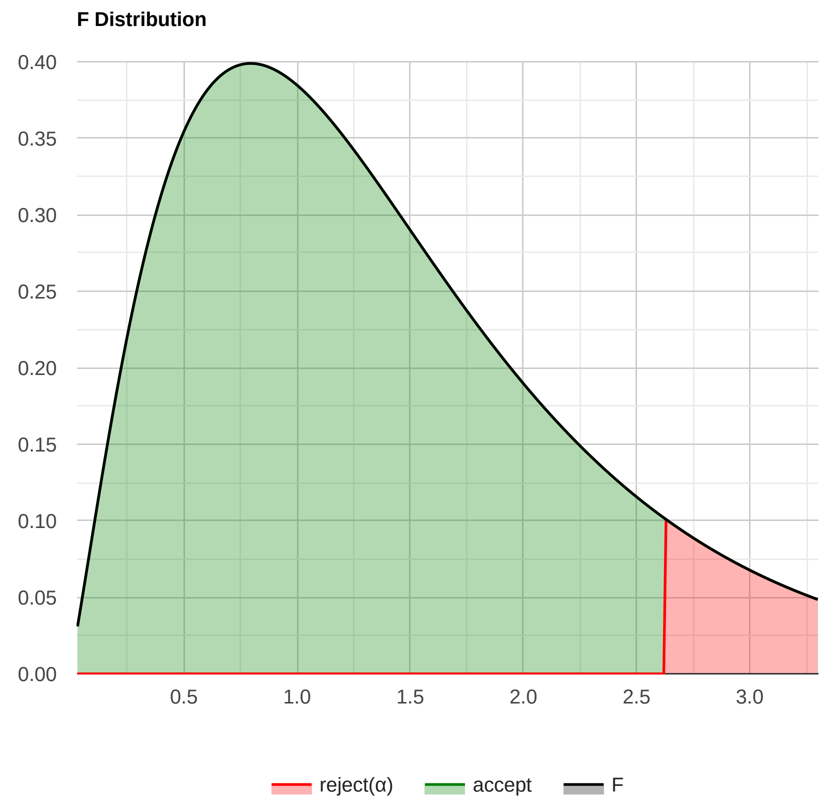
- **Residual normality**
  linear regression assumes normality for residual errors. Shapiro Wilk p-value equals 0. It is assumed that the data is not normally distributed.

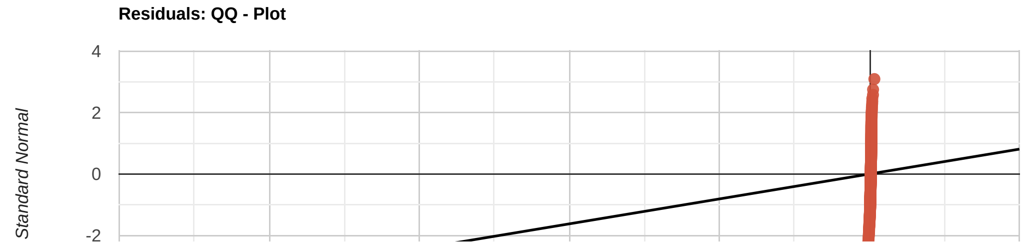
- **Homoscedasticity - homogeneity of variance**
  The White test p-value equals 0.722577 (F=0.325143). It is assumed that the variance is homogeneous.

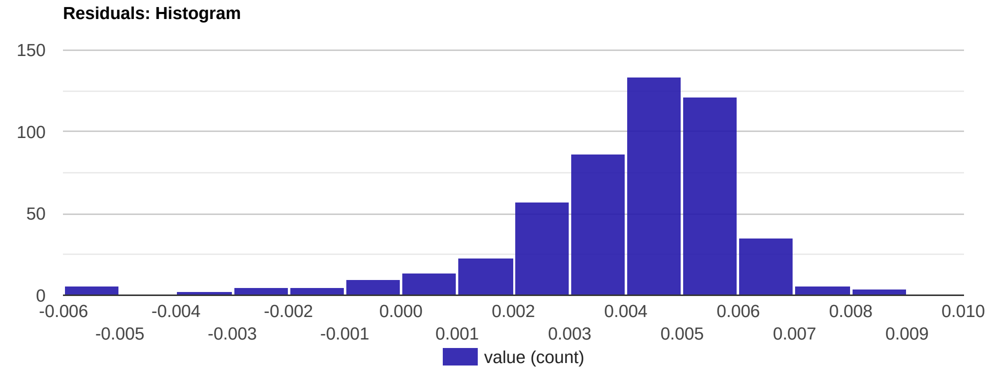
- **Multicollinearity - intercorrelations among the predictors (X_i)**
  There is a **moderate** multicollinearity concern as some of the VIF values are bigger than 5
  The multicollinearity may influence the coefficients or the ability to choose the predictors, but not the dependent variable (Y). There is no clear cut what is the VIF threshold, you should start being concerned for a value above 2.5. A value above 5 or 10 is probably not acceptable.

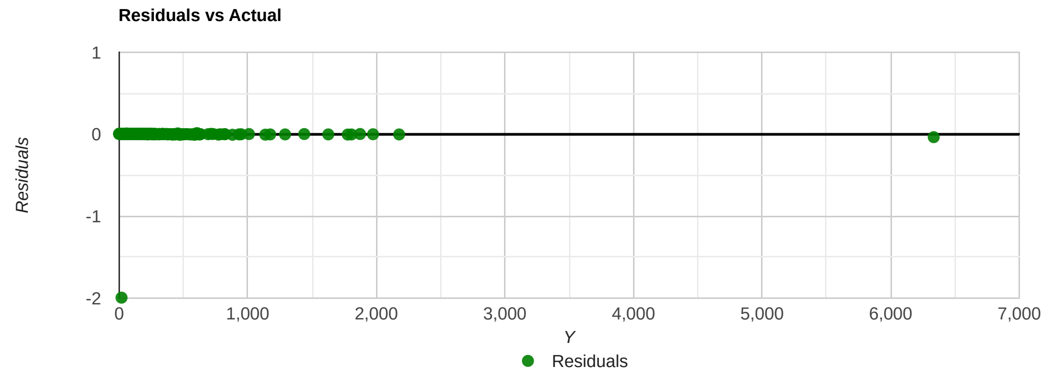
  You should remove **X_3** from the model unless the high multicollinearity cause is:
  1. Predictor combination (Like X_1X_2 or X_1^2.)
  2. Control variable, but the non-control variables do not have high multicollinearity.
  3. Dummy variable with more than two categories, when the reference category's proportion is small (the category that doesn't get a dummy variable).

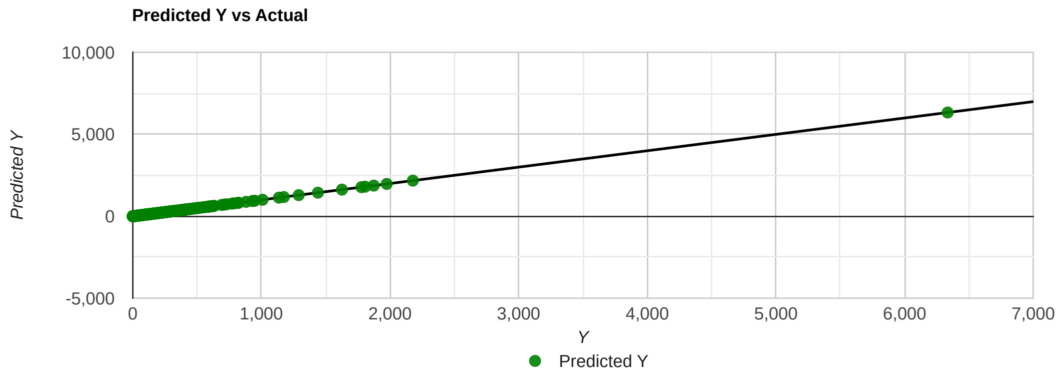
- **Priori power - of the entire model (6 predictors)**
  The **priori** power should be calculated **before** running the regression.
  The power to test the entire model is **strong**: **1**The power to prove that each predictor is significant is always lower than the power to test the entire model.

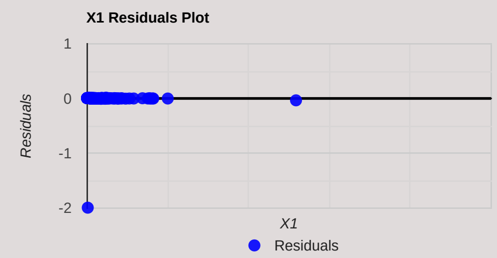


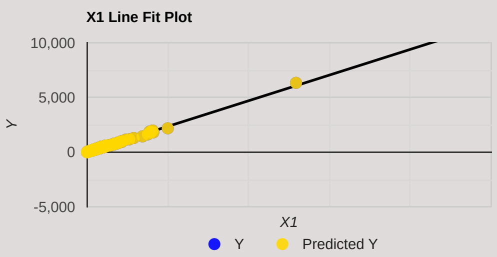

Residuals: QQ - Plot
Standard Normal
4
2
0
-2

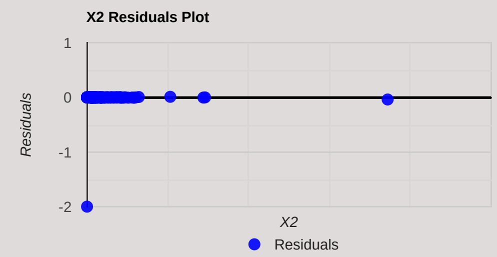

Residuals: Histogram
150
100
50
0
-0.006
-0.005
-0.004
-0.003
-0.002
-0.001
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.009
0.010
value (count)

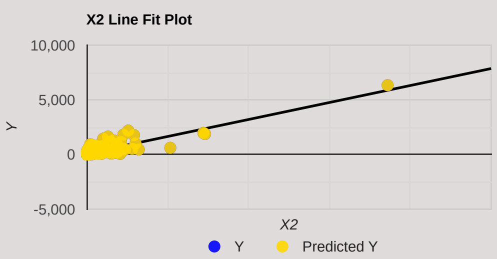

Residuals vs Actual
Residuals
1
0
-1
-2
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
Y
Residuals

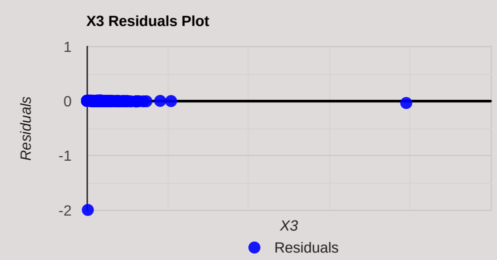

Predicted Y vs Actual
Predicted Y
10,000
5,000
0
-5,000
0
1,000
2,000
3,000
4,000
5,000
6,000
7,000
Y
Predicted Y

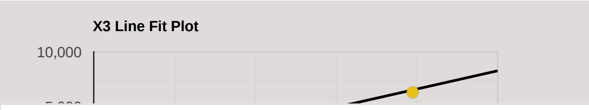

X1 Residuals Plot
Residuals
1
0
-1
-2
X1
Residuals


X1 Line Fit Plot
Y
10,000
5,000
0
-5,000
X1
Y
Predicted Y


X2 Residuals Plot
Residuals
1
0
-1
-2
X2
Residuals


X2 Line Fit Plot
Y
10,000
5,000
0
-5,000
X2
Y
Predicted Y


X3 Residuals Plot
Residuals
1
0
-1
-2
X3
Residuals



Multiple linear regression calculator

The calculator uses variables transformations, calculates the Linear equation, R, p-value, outliers and the adjusted Fisher-Pearson coefficient of skewness.
After checking the residuals' normality, multicollinearity, homoscedasticity and priori power, the program interprets the results.
Then, it draws a histogram, a residuals QQ-plot, a correlation matrix, a residuals x-plot and a distribution chart.
You may transform the variables, exclude any predictor or run backward stepwise selection automatically based on the predictor's p-value.

Right-tailed F test. Checks if the entire regression model is statistically significant. Why?

Multiple linear regression formula

$$\mathbf{Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_pX_p + \varepsilon}$$

It is easier to use the matrix form for multiple linear regression calculations:

$$\mathbf{Y = XB + E}$$

$$\hat{Y} = XB$$

$$B = (X'X)^{-1}X'Y$$

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ 1 & X_{31} & X_{32} & \dots & X_{3p} \\ 1 & X_{41} & X_{42} & \dots & X_{4p} \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \quad Y = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_n \end{bmatrix} \quad E = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_n \end{bmatrix}$$

$$B = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \\ \dots \\ B_p \end{bmatrix}$$

Y - dependent variable vector.
$\hat{Y}$ - predicted Y vector.
E - residuals vector, $E = Y - \hat{Y}$.
p - number of predictors.
n - sample size.

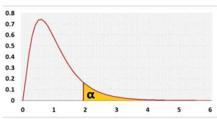
Hypotheses

H_0: $Y = b_0$
H_1: $Y = b_0 + b_1X_1 + ... + b_pX_p$

Test statistic

$$F = \frac{MS(Reg)}{MS(Res)}$$

F distribution



R Code

The following R code should produce the same results:

```
if(!"car" %in% installed.packages()){install.packages("car")}
library("car")
y0 <-
c(0,0,0,0,0,0,2,2,2,3,3,4,4,5,5,6,6,6,7,7,7,7,8,9,9,10,10,10,10,12,14,14,15,15,17,17,17,17,18,18,18,21,24,25,25,28,28,29,30,32,32,34,34,34,35,36,36,38,38,38,39,39,39,39,39,39,40,41,42,42,44,44,45,45,46,48,49,4
y1 <-
c(163,163,163,163,164,164,165,165,166,166,167,168,170,171,172,172,173,174,174,176,176,178,179,179,180,180,182,183,184,185,186,186,188,188,188,189,189,190,190,192,193,194,194,195,199,199,200,200,
y <- c(y0,y1)
x10 <-
c(0,0,0,0,0,0,2,2,2,3,3,1,4,4,4,4,3,6,7,7,6,5,6,7,7,9,8,7,7,11,12,12,11,14,13,15,13,15,16,18,15,14,19,23,22,17,24,22,25,18,30,28,28,25,24,26,28,29,30,32,27,30,32,34,27,32,35,36,34,39,37,36,40,33,29,32,40,32,36
x11 <-
c(130,117,142,129,130,146,152,128,139,129,148,138,128,143,141,148,142,143,144,134,125,154,148,156,139,142,136,140,154,153,152,145,154,148,138,165,138,155,155,161,161,154,148,162,172,150,166,163,
x1 <- c(x10,x11)
x20 <-
c(0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,0,1,0,0,0,3,1,0,0,0,6,0,0,0,2,0,1,0,6,0,2,2,3,0,7,1,2,2,0,1,3,4,1,1,0,0,0,0,0,4,1,0,2,7,0,1,2,3,0,3,2,0,2,1,4,0,3,7,0,41,0,1,1,1,2,1,0,0,0,3,1,0,3,2,1,5,1,0,5,2,2
x21 <-
c(2,16,3,2,5,3,2,5,2,1,3,4,4,1,5,3,7,18,1,8,10,6,3,3,7,1,4,5,4,5,1,10,4,12,12,2,12,12,3,0,7,4,7,26,6,20,2,2,10,15,9,10,4,2,0,3,1,6,6,8,4,3,3,5,11,9,2,9,7,6,9,2,3,38,10,4,6,5,14,6,9,18,9,12,17,2,1,9,3,5,4,4,7,20,0,5,30,6
x2 <- c(x20,x21)
x30 <-
c(0,0,0,0,0,0,0,0,0,0,0,2,0,1,1,2,2,0,0,0,1,2,2,2,2,1,2,3,2,1,1,2,3,1,4,2,1,1,2,2,3,1,5,2,3,9,4,6,5,8,2,4,4,6,11,3,7,7,6,6,11,6,3,4,11,7,5,5,8,3,3,7,5,10,10,16,8,15,11,13,9,4,14,12,10,9,7,10,11,8,1,11,9,6,14,6,7,10,15,14
x31 <-
c(31,30,18,32,29,15,11,32,25,36,16,26,38,27,26,21,24,13,29,34,41,18,28,20,34,37,42,38,26,27,33,31,30,28,38,22,39,23,32,31,25,36,39,7,21,29,32,35,13,44,41,26,33,47,15,29,30,18,27,18,36,26,26,53,5,23,54,32,
x3 <- c(x30,x31)
x40 <-
c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,3,1,1,1,2,1,1,1,1,2,1,1,1,2,1,1,3,1,1,1,1,1,1,1,1,1,1,1,2,1,1,1,1,1,1,1,3,1,1,1,1,3,1,1,2,2,2,1,1,1,1,3,1,1,1,3,1,1,2,1,2,1,1,3,2,3,1,3,1,1,2,1,3,3,1,3,2,1,2,3,
x41 <-
c(1,2,3,2,2,2,1,3,2,2,3,2,3,3,2,1,2,1,2,3,3,1,2,3,3,2,3,3,3,3,2,2,1,2,1,2,3,3,2,2,1,2,1,1,2,2,2,2,1,1,3,2,3,3,3,1,3,1,3,1,1,3,3,3,1,3,1,3,3,1,3,3,3,2,1,1,2,2,1,1,1,2,1,1,2,3,2,1,3,2,3,1,1,3,2,1,2,3,1,3,2,2,3,3,1,2,1,2,3,3,1,3,
x4 <- c(x40,x41)
x50 <-
c(10,2,9,8,10,4,9,19,15,12,15,8,12,13,17,12,7,14,11,6,18,4,10,10,5,13,11,9,9,8,2,2,11,4,4,13,10,15,10,10,11,6,11,7,13,4,12,9,4,2,9,10,6,15,11,10,10,12,2,18,10,3,8,4,10,3,8,11,11,12,13,4,9,3,3,10,12,6,1,3,3,9,4,3,1
x51 <-
c(3,11,13,3,10,10,6,13,3,7,13,13,3,12,3,6,3,11,14,2,10,6,4,13,4,2,4,7,3,10,9,11,13,13,8,10,2,8,9,3,10,10,3,3,11,4,11,3,10,3,13,1,13,3,3,14,15,3,10,14,11,9,8,13,9,4,11,11,3,4,13,13,10,10,2,17,9,11,2,11,13,5,13,17,4,
x5 <- c(x50,x51)
x60 <-
c(12,11,10,10,3,3,10,3,3,10,3,10,3,10,3,10,10,3,10,10,3,1,10,10,10,10,10,10,10,10,12,1,8,6,10,10,7,3,12,9,6,7,6,6,6,8,6,2,7,12,8,10,10,3,10,12,3,11,9,3,12,11,10,6,3,3,4,6,12,8,4,4,8,6,12,10,6,3,8,11,10,7,3,12,7,4,
x61 <-
c(11,9,6,2,12,9,4,2,8,5,5,4,2,7,6,2,4,12,5,2,1,2,10,5,5,5,2,4,6,10,12,7,7,6,5,9,4,5,5,12,10,9,6,7,12,9,10,4,11,10,7,7,5,12,10,3,4,12,8,4,12,9,7,2,6,6,11,10,2,9,4,6,1,9,3,3,10,5,11,7,6,6,4,3,10,6,2,11,10,9,8,11,3,5,12,1
x6 <- c(x60,x61)
model1 = lm(y~x1+x2+x3+x4+x5+x6)
summary(model1)
vif(model1)
```