CORRELATION AND REGRESSION ANALYSIS FOR OXYGEN CONSUMPTION PREDICTION

ZHELAL KHATIKE
Suleyman Demirel University

150107080@stu.sdu.edu.kz

Advisor: Aipenova Aziza Suleyman Demirel University, *PhD Computer Science*

INTRODUCTION

In this thesis, we will practice correlation and regression analysis for oxygen consumption with the help of other variables, such as: Run Time, Age, Weight, Run Pulse, Rest Pulse, Maximum Pulse and Performance. Also we will predict Oxygen Consumption using given variables.

Correlation statistics measure the strength of the linear relationship between two continuous variables. Two variables are correlated if there is a **linear** association between them. A common correlation statistic used for continuous variables is the Pearson correlation coefficient, denoted **r**, ranges between -1 and +1 and **quantifies the direction and strength of the linear association** between the two variables.

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Regression analysis is a widely used technique which is useful for evaluating multiple independent variables. Multiple linear regression analysis is an extension of simple linear regression analysis, used to assess the association between two or more independent variables and a single continuous dependent variable. The multiple linear regression equation is as follows:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p \text{ ,}$$

Y – the dependent variable

a – the intercept, also can be represented as b0

$b_1 – b_p$ – the slope of the line

X1, X2, Xn – used for multiple regression

The *a* and *b* can be found by following formulas:

$$A = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$B = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

The **coefficient of determination**, denoted R^2 or $r^2$and pronounced "R squared", is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

- The total sum of squares (proportional to the variance of the data), the analyst takes the predicted Y value and subtracts the average actual value:

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

- The sum of squares of residuals, also called the residual sum of squares:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

- The most general definition of the coefficient of determination is

$$R^2 \equiv 1 - \frac{SS_{\rm res}}{SS_{\rm tot}}$$

DATA

The data we used is **Fitness** table, with 31 observations and 6 values as: Oxygen Consumption, Run Time, Age, Weight, Run Pulse, Rest Pulse, Maximum Pulse and Performance.

Output Results

Obs	Name	Gender	RunTime	Age	Weight	Oxygen_Consumption	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
1	Donna	F	8.17	42	68.15	59.57	166	40	172	90
2	Gracie	F	8.63	38	81.87	60.06	170	48	186	94
3	Luanne	F	8.65	43	85.84	54.30	156	45	168	83
4	Mimi	F	8.92	50	70.87	54.63	146	48	155	67
5	Chris	M	8.95	49	81.42	49.16	180	44	185	72
6	Allen	M	9.22	38	89.02	49.87	178	55	180	92
7	Nancy	F	9.40	49	76.32	48.67	186	56	188	64
8	Patty	F	9.63	52	76.32	45.44	164	48	166	56
9	Suzanne	F	9.93	57	59.08	50.55	148	49	155	43
10	Teresa	F	10.00	51	77.91	46.67	162	48	168	54
11	Bob	M	10.07	40	75.07	45.31	185	62	185	79
12	Harriett	F	10.08	49	73.37	50.39	168	67	168	57
13	Jane	F	10.13	44	73.03	50.54	168	45	168	67
14	Harold	M	10.25	48	91.63	46.77	162	48	164	61
15	Sammy	M	10.33	54	83.12	51.85	166	50	170	49
16	Buffy	F	10.47	52	73.71	45.79	186	59	188	47
17	Trent	M	10.50	52	82.78	47.47	170	53	172	51
18	Jackie	F	10.60	47	79.15	47.27	162	47	164	56
19	Ralph	M	10.85	43	81.19	49.09	162	64	170	65
20	Jack	M	10.95	51	69.63	40.84	168	57	172	48
21	Annie	F	11.08	51	67.25	45.12	172	48	172	43
22	Kate	F	11.12	45	66.45	44.75	176	51	176	55
23	Carl	M	11.17	54	79.38	46.08	156	62	165	40
24	Don	M	11.37	44	89.47	44.61	178	62	182	58
25	Effie	F	11.50	48	61.24	47.92	170	52	176	45
26	George	M	11.63	47	77.45	44.81	176	58	176	50
27	Iris	F	11.95	40	75.98	45.68	176	70	180	56
28	Mark	M	12.63	57	73.37	39.41	174	58	176	20
29	Steve	M	12.88	54	91.63	39.20	168	44	172	23
30	Vaughn	M	13.08	44	81.42	39.44	174	63	176	41
31	William	M	14.03	45	87.66	37.39	186	56	192	30

CORRELATION ANALYSIS

In this program, it’s possible to take a real data and see correlation of ***oxygen consumption*** with all variables on SAS:

```
proc corr data=statdata.fitness rank

     plots(only)=scatter(nvar=all ellipse=none);

   var RunTime Age Weight Run_Pulse

       Rest_Pulse Maximum_Pulse Performance;

   with Oxygen_Consumption;

   title "Correlations and Scatter Plots with Oxygen_Consumption";

run;

title;
```

Correlations and Scatter Plots with Oxygen_Consumption

The CORR Procedure

1 With Variables:	Oxygen_Consumption
7 Variables:	RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
Oxygen_Consumption	31	47.37581	5.32777	1469	37.39000	60.06000
RunTime	31	10.58613	1.38741	328.17000	8.17000	14.03000
Age	31	47.67742	5.26236	1478	38.00000	57.00000
Weight	31	77.44452	8.32857	2401	59.08000	91.63000
Run_Pulse	31	169.64516	10.25199	5259	146.00000	186.00000
Rest_Pulse	31	53.45161	7.61944	1657	40.00000	70.00000
Maximum_Pulse	31	173.77419	9.16410	5387	155.00000	192.00000
Performance	31	56.64516	18.32584	1756	20.00000	94.00000

Pearson Correlation Coefficients, N = 31 Prob > |r| under H0: Rho=0							
Oxygen_Consumption	RunTime -0.86219 <.0001	Performance 0.77890 <.0001	Rest_Pulse -0.39935 0.0260	Run_Pulse -0.39808 0.0266	Age -0.31162 0.0879	Maximum_Pulse -0.23677 0.1997	Weight -0.16289 0.3813

In this table we see Oxygen_Consumption variable is used as x and others as y.

In the following program, it’s possible to see correlation of all variables with the all variables as a matrix, it’s means we take all variables as x and all variables as y too , on SAS:

```
ods graphics on / imagemap=on;

proc corr data=statdata.fitness nosimple

     plots=matrix(nvar=all histogram);

   var RunTime Age Weight Run_Pulse

       Rest_Pulse Maximum_Pulse Performance;

   id name;
```

```
title "Correlation Matrix and Scatter Plot Matrix of Fitness Predictors";

run;

title;
```

Correlation Matrix and Scatter Plot Matrix of Fitness Predictors

The CORR Procedure

8 Variables:	Oxygen_Consumption RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

Pearson Correlation Coefficients, N = 31
Prob > |r| under H0: Rho=0

	Oxygen_Consumption	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
Oxygen_Consumption	1.00000	-0.86219 <.0001	-0.31162 0.0879	-0.16289 0.3813	-0.39808 0.0266	-0.39935 0.0260	-0.23677 0.1997	0.77890 <.0001
RunTime	-0.86219 <.0001	1.00000	0.19523 0.2926	0.14351 0.4412	0.31365 0.0858	0.45038 0.0110	0.22610 0.2213	-0.82049 <.0001
Age	-0.31162 0.0879	0.19523 0.2926	1.00000	-0.24050 0.1925	-0.31607 0.0832	-0.15087 0.4178	-0.41490 0.0203	-0.71257 <.0001
Weight	-0.16289 0.3813	0.14351 0.4412	-0.24050 0.1925	1.00000	0.18152 0.3284	0.04397 0.8143	0.24938 0.1761	0.08974 0.6312
Run_Pulse	-0.39808 0.0266	0.31365 0.0858	-0.31607 0.0832	0.18152 0.3284	1.00000	0.35246 0.0518	0.92975 <.0001	-0.02943 0.8751
Rest_Pulse	-0.39935 0.0260	0.45038 0.0110	-0.15087 0.4178	0.04397 0.8143	0.35246 0.0518	1.00000	0.30512 0.0951	-0.22560 0.2224
Maximum_Pulse	-0.23677 0.1997	0.22610 0.2213	-0.41490 0.0203	0.24938 0.1761	0.92975 <.0001	0.30512 0.0951	1.00000	0.09002 0.6301
Performance	0.77890 <.0001	-0.82049 <.0001	-0.71257 <.0001	0.08974 0.6312	-0.02943 0.8751	-0.22560 0.2224	0.09002 0.6301	1.00000

This table represents all correlations between variables.

We can see that Oxygen Consumption has **strong positive correlation** with performance (noticed by green) because it has more than 0.5 correlation coefficient; Age, Run Pulse and Rest Pulse have **weak negative** (orange), Weight and Minimum Pulse have **about to no correlation** (yellow), and oxygen Consumption with Run Time **strong negative correlation** (red).

MULTIPLE REGRESSION ANALYSIS

In this program, let's take a look an example of multiple regression using fitness data set on SAS:

```
proc reg data=statdata.fitness;
  model Oxygen_Consumption = RunTime Age Weight Run_Pulse
        Rest_Pulse Maximum_Pulse Performance;
  title 'Regression of Fitness on All Predictors';
run;
quit;
title;
```

Regression of Fitness on All Predictors

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen_Consumption

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.66124	103.23732	18.42	<.0001
Error	23	128.89331	5.60406		
Corrected Total	30	851.55455			

Root MSE	2.36729	R-Square	0.8486
Dependent Mean	47.37581	Adj R-Sq	0.8026
Coeff Var	4.99683		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	131.78249	72.20754	1.83	0.0810
RunTime	1	-3.86019	2.93659	-1.31	0.2016
Age	1	-0.46082	0.58660	-0.79	0.4401
Weight	1	-0.05812	0.06892	-0.84	0.4078
Run_Pulse	1	-0.36207	0.12324	-2.94	0.0074
Rest_Pulse	1	-0.01512	0.06817	-0.22	0.8264
Maximum_Pulse	1	0.30102	0.13981	2.15	0.0420
Performance	1	-0.12619	0.30097	-0.42	0.6789

In our case, **R-Squared** (blue) is total 85%.– It's mean our predicted values are true as like table for 85 percent.

Intercept – is our a, and it is – 131.78249

The slope of the line – or also it is X1, X2, X3, X4, X5, X6, X7 **.**

Equation will be represented as follows: $Y = 131.78 - 3.86*x_{runtime} - 0.46*x_{age} - 0.058*x_{weight} - 0.36*x_{runpulse} - 0.015*x_{restpulse} + 0.3*x_{maxpulse} - 0.126*x_{performance}$

PREDICT OXYGEN CONSUMPTION WITH THE HELP OF OTHER VARIABLES

In this program, we can predict oxygen consumption with the help of other variables predictors using fitness data set on SAS:

```
data need_predictions;
   infile datalines dlm='/';
   input RunTime Age Weight Run_Pulse Rest_Pulse
Maximum_Pulse Performance;
   datalines;
8.17/42/68.15/166/40/172/90
8.63/38/81.87/170/48/186/94
10.00/51/60/140/60/180/80
;
run;
data predoxy;
   set need_predictions
       statdata.fitness;
run;
proc reg data=predoxy;
   model Oxygen_Consumption=RunTime Age Weight Run_Pulse
```

```
        Rest_Pulse Maximum_Pulse Performance / p;
    id RunTime Age Weight Run_Pulse
        Rest_Pulse Maximum_Pulse Performance;
    title 'Oxygen_Consumption=RunTime with Predicted Values';
run;
quit;
title;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen_Consumption

Output Statistics										
Obs	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance	Dependent Variable	Predicted Value	Residual
1	8.17	42	68.2	166	40	172	90	.	56.6390	.
2	8.63	38	81.9	170	48	186	94	.	58.0494	.
3	10.00	51	60.0	140	60	180	80	.	58.6825	.
4	8.17	42	68.2	166	40	172	90	59.6	56.6390	2.9310
5	8.63	38	81.9	170	48	186	94	60.1	58.0494	2.0106
6	8.65	43	85.8	156	45	168	83	54.3	56.5215	-2.2215
7	8.92	50	70.9	146	48	155	67	54.6	54.8047	-0.1747
8	8.95	49	81.4	180	44	185	72	49.2	50.6863	-1.5263
9	9.22	38	89.0	178	55	180	92	49.9	50.8002	-0.9302
10	9.40	49	76.3	186	56	188	64	48.7	48.8044	-0.1344
11	9.63	52	76.3	164	48	166	56	45.4	49.0077	-3.5677
12	9.93	57	59.1	148	49	155	43	50.6	50.6548	-0.1048
13	10.00	51	77.9	162	48	168	54	46.7	49.5264	-2.8564
14	10.07	40	75.1	185	62	185	79	45.3	47.9135	-2.6035
15	10.08	49	73.4	168	67	168	57	50.4	47.5647	2.8253
16	10.13	44	73.0	168	45	168	67	50.5	48.7664	1.7736
17	10.25	48	91.6	162	48	164	61	46.8	47.0590	-0.2890
18	10.33	54	83.1	166	50	170	49	51.9	46.3217	5.5283
19	10.47	52	73.7	186	59	188	47	45.8	45.5431	0.2469
20	10.50	52	82.8	170	53	172	51	47.5	45.4629	2.0071

Here we can see **input values** (green) and **predicted value** (red).

Equation will be represented as follows: Y= 131.78 – (3.86*10) – (0.46*51) - (0.058*60) – (0.36*140) – (0.015*60) + (0.3*180) – (0.126*80) **= 58.6825**

58.6825 – the predicted oxygen consumption for specific input values, calculated by SAS.

CONCLUSION

In this thesis, we finalize that using correlation and regression analysis, we can analyze relationship between values. Correlation statistics measure the strength of the linear relationship between two continuous variables in numerical form. Two variables are correlated if there is a linear association between them, denoted **r**, ranges between -1 and +1 and **quantifies the direction and strength of the linear association** between the two variables.

Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). It is useful for evaluating multiple independent variables. Multiple linear regression analysis is an extension of simple linear regression analysis, used to assess the association between two or more independent variables and a single continuous dependent variable.

During the analysis, we saw how the correlation and multiple regression is works, using fitness data set observations examine relationship of their values. Find out why Oxygen Consumption of person and it’s performance has strong positive dependence and others has negative dependence.