Report on Linear Regression

Course: 837 Machine Learning

Group: VIII

Submitted by
Amran Hossain (BSSE 0917)
Niraj Chaudhary (BSSE 0942)
Zahid Hasan Rifad (BSSE 0944)

Submitted To
Dr. B M Mainul Hossain
Associate Professor
Institute of Information Technology
University of Dhaka

Institute of Information Technology

University of Dhaka

Date: 9 Jan, 2021

Table of Contents

Linear Regression

Linear regression is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables)

Dataset Description

We have chosen a dataset which contain diabetes type1 data and pick BMI related data for fitting the linear regression. There are 307 rows and 22 columns in the main dataset. For our purpose we have chosen 6 columns. These columns belongs to height, weight, BMI, gender, area and age. Also, there are 3 categorical variables such as gender, area, age. We have taken BMI as a y variable from our dataset.

Model

We have built many regression models from our dataset. There are four models in details.

Model-1

Title:

$$BMI = \beta 0 + \beta 1 * \left(\frac{1}{weight}\right) + \beta 2 * (height)^2 + \beta 3 * \ln(age) + \beta 4 * \text{area}$$

Here

Y= BMI

X = $\frac{1}{weight}, (height)^2, \ln(age), area$

R^2 calculation:

The R^2 value of this model is **12%**

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.349991866							
5	R Square	0.122494306							
6	Adjusted R Square	0.110833101							
7	Standard Error	5.870606442							
8	Observations	306							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	4	1448.099448	362.0248619	10.50442931	5.59384E-08			
13	Residual	301	10373.67002	34.46402					
14	Total	305	11821.76947						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	29.62434752	1.757068276	16.86010038	2.26741E-45	26.16665409	33.0820409	26.16665409	33.08204094
18	X Variable 1	-57.09933751	12.08231535	-4.725860554	3.52267E-06	-80.87584233	-33.322833	-80.8758423	-33.32283269
19	X Variable 2	-4.614134177	0.880407942	-5.240904763	3.01211E-07	-6.346668304	-2.8816	-6.3466683	-2.88160005
20	X Variable 3	2.627967701	1.455247527	1.805856153	0.071939704	-0.235779755	5.49171516	-0.23577976	5.491715157
21	X Variable 4	-0.302497914	0.453174377	-0.667508865	0.504958584	-1.194289134	0.58929331	-1.19428913	0.589293307
22									

Figure 1: result of model 1

Model-2:

Title:

$$\ln(BMI) = \beta 0 + \beta 1 * (weight) + \beta 2 * (height)^2 + \beta 3 * age + \beta 4 * \ln(age)\text{^}2$$

Here

Y= $\ln(BMI)$

X = $weight$, $(height)^2$, age and $\ln(age)\text{^}2$

Difference between previous model:

Previous model y variable was BMI but, in this model, we have used e base log of BMI. Independent variable we used (height)^2, weight, age and ln(age)^2 but previous model we used 1/weight, height^2, ln(age), area. For this reason, R^2 value has been increased.

R^2 calculation:

The R^2 value of this model is **32%**

A1 SUMMARY OUTPUT

	A	B	C	D	E	F	G	H	I	J
1	SUMMARY OUTPUT									
2										
3	*Regression Statistics*									
4	Multiple R	0.572017553								
5	R Square	0.327204081								
6	Adjusted R Square	0.318263272								
7	Standard Error	0.218025673								
8	Observations	306								
9										
10	ANOVA									
11		*df*	*SS*	*MS*	*F*	*Significance F*				
12	Regression	4	6.958523995	1.739631	36.5967	6.27379E-25				
13	Residual	301	14.30809337	0.047535						
14	Total	305	21.26661736							
15										
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*pper 95.0%*	
17	Intercept	3.289723412	0.551450874	5.965578	6.83E-09	2.20453618	4.374911	2.204536	4.374911	
18	X Variable 1	0.013008079	0.001136713	11.44359	2.04E-25	0.010771167	0.015245	0.010771	0.015245	
19	X Variable 2	0.059188399	0.066548423	0.889403	0.374497	-0.071770681	0.190147	-0.07177	0.190147	
20	X Variable 3	-0.455562801	0.630272593	-0.7228	0.470362	-1.695861445	0.784736	-1.69586	0.784736	
21	X Variable 4	-0.032842113	0.660920274	-0.04969	0.960401	-1.333451609	1.267767	-1.33345	1.267767	
22										
23										
24										
25	RESIDUAL OUTPUT									

Figure-2: result of model 2

Model-3:

Title:

$$\text{BMI} = \beta 0 + \beta 1 * (height)^2 + \beta 2 * (weight)$$

Here

Y= BMI

X = $weight$, $(height)^2$

Difference between previous model:

Previous model y variable was log of BMI but in this model, we have used normal BMI. Independent variable we used (height)^2 and weight but previous model we used age and log of age^2. For this reason, R^2 value has been increased.

R^2 calculation:

The R^2 value of this model is **69%**

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.833619844
R Square	0.694922044
Adjusted R Square	0.692908328
Standard Error	3.450049634
Observations	306

ANOVA

	df	SS	MS	F	Significance F
Regression	2	8215.208197	4107.604098	345.0943845	7.73109E-79
Residual	303	3606.561271	11.90284248		
Total	305	11821.76947			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	24.08163069	0.587331781	41.00174974	1.1293E-125	22.92586506	25.23739632	22.92586506	25.23739632
X Variable 1	-12.99295818	0.530820664	-24.47711449	9.21226E-74	-14.03751987	-11.94839648	-14.03751987	-11.94839648
X Variable 2	0.552036817	0.021483272	25.69612402	4.38058E-78	0.509761517	0.594312117	0.509761517	0.594312117

Figure-3: result of model 3

Model-4:

Title:

$$\ln(BMI) = \beta 0 + \beta 1 * ln(height) + \beta 2 * ln(weight) + \beta 3 * \ln(age) + \beta 4 * \ln(gender) + \beta 5 * \ln(Residence)$$

Here

Y= $\ln(BMI)$

X = $height, weight, age, gender, Residence$

Difference between previous model:

Previous model y variable was normal BMI but, in this model, we have used e base log of BMI. Independent variable we used height^2 and weight in previous model but in this model, we used e base log of all independent variables. E base log of (height, weight, gender, age, Residence). R^2 value has been increased.

R^2 calculation:

The R^2 value of this model is **89.8%**

A	B	C	D	E	F	G	H	I
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.947861983							
R Square	0.898442338							
Adjusted R Square	0.89674971							
Standard Error	0.084848648							
Observations	306							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	19.10683	3.821366	530.797375	1.2121E-146			
Residual	300	2.159788	0.007199					
Total	305	21.26662						
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Upper 95.0%*
Intercept	0.46531909	0.053595	8.68211	2.5338E-16	0.359849014	0.5707892	0.359849	0.57078917
X Variable 1	-1.774579697	0.040598	-43.7107	3.85E-132	-1.854473279	-1.6946861	-1.85447	-1.69468612
X Variable 2	0.847787613	0.018169	46.6623	1.447E-139	0.812033613	0.8835416	0.812034	0.88354161
X Variable 3	0.012480805	0.021512	0.580183	0.56222643	-0.029852383	0.054814	-0.02985	0.05481399
X Variable 4	0.026857874	0.014184	1.893544	0.05924764	-0.001054706	0.0547705	-0.00105	0.05477045
X Variable 5	-0.005588036	0.01169	-0.47802	0.63298649	-0.028592909	0.0174168	-0.02859	0.01741684

Figure-4: result of model 4