



Gen. Math. Notes, Vol. 21, No. 1, March 2014, pp. 118-127

ISSN 2219-7184; Copyright © ICSRS Publication, 2014

www.i-csrs.org

Available free online at <http://www.geman.in>

Comparison between Models With and Without Intercept

Sameera Abdulsalam Othman

Department of Mathematics, Faculty of Educational Science
School of Basic Education, University of Duhok, Duhok-Iraq
E-mail: samira.a@hotmail.com

(Received: 8-1-14 / Accepted: 25-2-14)

Abstract

The aim of this paper is Comparison between models with and without intercept and Statement the beast one, and applying the method leverage point when we added the new point to the original data. We are testing the significant intercept by using (t) test.

Keywords: *Intercept, Hypothesis, Significant, Original, Regression.*

1.1 Introduction

Multiple linear regressions (MLR) is a method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is sometimes also called the predict and the independent variables the predictors. The aim of this paper is Comparison between models with and without intercept and Statement the beast one, and contain the important definition of the regression and the most important relationship and the equation that are used to solve example about the Multiple linear regression of least squares and estimation and test of hypothesis due to the parameters, and so the most

important application the theoretical of blood pressure (dependent variable Y) and height, weight, age, sugar, sex, hereditary factor social status (independent variable).

2.1 Linear Regression Models and Its Types

a. Linear Regression Model with Intercept

The linear regression be intercept if the line regression intersection with Y axis in not origin. It means that mathematically $B_0 \neq 0$ that is intersection point of regression line with Y axis

$$Y_i = B_0 + B_1 X_{i1} + e_i \quad , i = 1, 2, 3, \dots, n \quad (1)$$

Y_i = depended variyable

X_{i1} = independent variable

B_0, B_1 = regression parameter

e_i = value of random error

b. Linear Regression Model without Intercept [4], [3]

The linear regression be without intercept when the line regression to pass through the origin. It means that mathematically $B_0 = 0$

We can write the simple linear regression model

$$Y_i = B_1 X_{i1} + e_i \quad (2)$$

The parameters B_0 and B_1 usually anknown and estimate by least squares method. From(3,4)

$$\hat{B}_1 = \frac{S_{xy}}{S_{xx}} \quad (3)$$

$$\hat{B}_0 = \bar{Y} - \hat{B}_1 \bar{X} \quad (4)$$

Where

S_{xy} : Standard deviation between X and Y

S_x : Standard deviation of X

But linear regression without intercept

$$\hat{B}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n x_i^2} \quad (5)$$

2.2 Leverage Point and Regression through the Origin Leverage Point

The interpretation can be use in understanding the difference between the full fit and the for forced through the origin. It is show that the regression through the origin is equivalent to fitting the full model. To a new data set. This new data a set is composed of the original observation. Evaluation of the leverage possessed by this new points is equivalent to evaluating when the $B_0 = 0$ in the full model.

2.3 Augmenting the Data Set

Let $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ be n data points observed according to $Y_i = b_0 + B_1 X_i + \epsilon_i$

Where the experimental errors, ϵ_i , are independently normally distributed with mean 0 and σ^2 . the least squares estimates for b_0 and b_1 are

$$\hat{b}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{b}_0 = \bar{y} - \hat{B}_1 \bar{X}$$

Where $\bar{X} = \frac{\sum X_i}{n}$, $\bar{Y} = \frac{\sum Y_i}{n}$ [1], [10]

If the regression is forced through the origin, then it is assumed that the data are observed according to

$Y_i = B_1 X_i + e_i$ and the least squares estimate of B_1 in this model is

$$\hat{B}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

If the original data set is augmented with a new observation

$$(X_{n+1}, Y_{n+1}) = (n^* \bar{X}, n^* \bar{Y})$$

Where $n^* = \frac{n}{\sqrt{n+1}-1}$ (6)

Then fitting the full model to the augmented data set is equivalent to forcing the original regression through the origin. This follows from the easily verified identities

$$\sum_{i=1}^{n+1} (X_i - \bar{X}_{n+1})(Y_i - \bar{Y}_{n+1}) = \sum_{i=1}^n X_i Y_i$$

$$\sum_{i=1}^{n+1} (X_i - \bar{X}_{n+1})^2 = \sum_{i=1}^n X_i^2 \quad , \quad \sum_{i=1}^{n+1} (Y_i - \bar{Y}_{n+1})^2 = \sum_{i=1}^n Y_i^2$$

Where

$$\bar{X}_{n+1} = \frac{\sum_{i=1}^{n+1} X_i}{n+1} \quad , \quad \bar{Y}_{n+1} = \frac{\sum_{i=1}^{n+1} Y_i}{n+1}$$

The position of the point $(n^*\bar{X}, n^*\bar{Y})$, relative to the other points, determines the new points has high or low leverage. The leverage of the new point can be used to decide if the regression through the origin is more a pirate intercept term. [5]

2.4 Assessing the Leverage of Data Point

The leverage, h_{ij} , of a data point Y_j is the amount of influence that a predicted value, say \hat{Y}_i , can be written as

$$\hat{Y}_i = \sum_{j=1}^n h_{ij} Y_j \quad \text{Where}$$

$$h_{ij} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_j - \bar{X})}{\sum_{k=1}^n (X_k - \bar{X})^2} \quad (7)$$

The h_{ij} show how each observation Y_j affects the predicted value Y_i . More importantly, however, h_{ii} show how Y_i affects \hat{Y}_i , and is quite use full in the detection of influential points. The relative size of h_{ii} can give us information on the potential influence Y_i has on the fit. For purposes of comparison, it is fortunate that the values h_{ii} have a built-in scale. The matrix $H = [h_{ij}]$ is a projection matrix and from the properties of projection matrices it can be verified that $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^n h_{ii} = p$, where p is the number of coefficients to be estimated. Thus, on the average, h_{ii} should be approximately equal p/n . Hoaglin and welsch suggest, as a rough guideline, paing attention to any data points having $h_{ii} > 3p/n$. The h_{ij} values depend only on the experimental design (the X^s), and not on the results of the results of the experiment (the Y^s), hence a data point with high leverage may not necessarily have an adverse effect on the fit. [7]

2.5 The Leverage of Augmented Point

We are concerned here with influence of the $(n+1)$ st data points $(n^*\bar{X}, n^*\bar{Y})$ using and it is stra l ght forward to calculate.

$$h_{n+1} = \frac{1}{n+1} \left[1 + \frac{n^2 \bar{X}^2}{\sum_{i=1}^n x_i^2} \right] \quad (8)$$

2.6 The Leverage:

There is also a straight forward generalization to multiple linear regression the original data is augmented with the point $(n^*\bar{X}, n^*\bar{Y})$ where \bar{X} is a vector containing the means of the independent variables. The impact of the augmented data point on the fit because cleared when h_{n+1} is also examined, it is, perhaps, more instructive to write h_{n+1} in the equivalent form.

$$h_{n+1} = \frac{1}{n+1} \left[1 + n \left[\frac{\bar{X}^2}{\sigma_x^2 + \bar{X}^2} \right] \right] \quad (9)$$

Where $\sigma_x^2 = \frac{\sum (X_i - \bar{X})^2}{n}$ (10)

Thus the impact of the augmented data point increases with $(\bar{X}/\sigma_x)^2$ and we can expect the greatest discrepancy between the full fit and through the origin when \bar{X} is large compared to σ_x^2 . The augmented data set will seem to be composed of two distinct cluster ; one composed of the original data and one composed of $(n^*\bar{X}, n^*\bar{Y})$. [2]

2.7 Analysis of Variance (ANOVA)

ANOVA enables us to draw inferences about whether the samples have been drawn from populations having the same mean. It is used to test for differences among the means of the populations by examining the amount of variation within each of the samples, relative to the amount of variation between the samples. In terms of variations within the given population, it is assumed that the values differ from the mean of this population only because of the random effects. The test statistics used for decision making is $F =$ ratio of Estimate of population variance based on between samples variance and Estimate of population variance based on within samples variance [8].

2.8 ANOVA Table

For a one-way ANOVA, the table looks like:

Source	df	SS	MS	F	p-value
Treatment	(k-1)	SST	MST	MST/MSE	p
Error	(N-k)	SSE	MSE		
Total	N-1	SST			[6], [4]

2.9 Testing of Hypotheses

We test the significant with and without intercept by using (t) test

$$H_0: \beta_0 = 0$$

$$H_1: \beta_0 \neq 0 \quad \text{and}$$

$$t_0 = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \quad (11)$$

Where $S(\hat{\beta}_0)$: stander deviation of $\hat{\beta}_0$ [9]

The result of testing the null hypothesis is reject if

$$|t_0| \geq t_{\left(\frac{\alpha}{2}, n-p-1\right)} \quad (12)$$

3.1 Application

This section contains an application of what the theoretic part mentioned in the section two. Data are obtained for the practical application of these samples from healthy centers. We take the data in the study of blood pressure (dependent variable Y) and height (X₁), weight (X₂), age (X₃) sugar (X₄), sex (X₅, 0 denote to male and 1 to female), hereditary factor (X₆, 0 if have hereditary factor, 1 if no), social status (X₇, 0 if married and 1 if single) (independent variable) for (100) person. We can show that in table (1).

Table (1): Blood pressure and height, weight, age, sugar, sex, hereditary factor for (100) person

No.	Y	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	No.	Y	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇
1	130	145	41	24	178	1	0	0	51	80	153	45	27	370	1	1	1
2	140	152	79	49	100	1	1	0	52	70	163	48	25	339	1	1	1
3	140	158	83	34	270	0	1	1	53	210	171	116	40	310	0	1	1
4	130	172	83	48	223	1	1	0	54	240	151	54	50	255	0	1	1
5	150	161	67	40	82	1	1	0	55	140	157	69	31	295	0	1	1
6	190	160	72	46	160	0	0	1	56	140	161	43	45	244	0	0	1
7	110	159	48	21	170	0	0	0	57	170	164	60	54	280	0	0	0
8	80	152	75	44	82	1	1	1	58	180	152	61	45	296	0	1	1
9	120	165	54	46	147	0	1	1	59	80	168	67	37	320	0	1	1
10	140	157	69	31	150	0	0	0	60	210	153	35	58	195	0	0	1
11	140	161	43	45	117	0	0	0	61	200	156	87	45	190	0	1	1
12	130	165	83	48	165	0	0	0	62	130	172	83	48	210	0	0	0
13	170	160	110	41	335	0	1	1	63	200	166	90	40	216	0	1	1
14	80	154	45	27	90	0	1	1	64	170	169	85	27	269	0	1	1
15	240	151	54	50	314	0	0	1	65	80	150	45	28	270	1	1	1

16	100	180	69	32	151	0	1	1	66	180	166	69	41	378	0	1	0
17	180	167	91	40	190	0	1	1	67	80	150	55	25	399	1	1	1
18	160	154	70	100	264	1	1	1	68	150	161	54	40	369	0	1	1
19	120	172	59	55	133	1	1	0	69	170	173	79	59	315	0	1	0
20	100	152	53	62	90	0	1	1	70	220	171	82	60	345	0	1	0
21	140	162	43	80	100	0	0	1	71	170	156	53	16	214	0	1	1
22	230	174	72	60	82	0	1	0	72	170	167	70	55	195	0	0	0
23	70	163	48	25	270	0	1	1	73	230	174	72	60	377	0	1	0
24	190	155	92	70	350	0	1	1	74	180	152	45	60	430	0	0	1
25	210	151	80	50	250	0	1	1	75	200	171	60	43	377	0	1	0
26	200	156	87	45	234	0	1	1	76	180	164	60	60	384	0	0	0
27	240	151	54	50	310	0	1	0	89	170	158	65	60	258	0	1	1
28	250	172	90	65	126	1	1	1	90	200	160	65	52	432	0	0	1
29	117	154	55	28	260	0	0	1	91	160	172	70	70	170	0	1	0
30	230	174	72	60	245	0	1	1	92	180	154	65	52	185	0	1	1
31	145	155	69	33	130	0	1	1	93	250	172	90	65	265	0	1	0
32	145	150	65	35	144	0	1	1	94	210	151	80	50	250	0	1	1
33	190	156	93	45	249	0	1	1	95	117	154	55	28	128	0	1	1
34	190	160	72	46	320	0	1	1	96	150	172	96	35	235	0	1	0
35	160	160	64	50	227	0	0	1	97	160	160	70	69	170	0	1	0
36	170	158	55	55	179	0	1	1	98	170	160	110	41	174	0	1	1
37	160	157	70	54	234	0	0	1	99	90	154	50	25	175	1	1	1
38	175	160	83	44	230	0	0	0	100	180	152	85	60	280	0	0	1
39	150	168	111	50	300	0	0	1	89	170	158	65	60	258	0	1	1
40	180	176	81	60	420	0	0	0	90	200	160	65	52	432	0	0	1
41	180	159	67	50	255	0	1	1	91	160	172	70	70	170	0	1	0
42	170	156	70	47	289	0	1	1	92	180	154	65	52	185	0	1	1
43	150	149	60	45	176	0	1	1	93	250	172	90	65	265	0	1	0
44	160	154	70	60	156	0	0	1	94	210	151	80	50	250	0	1	1
45	150	154	69	40	195	0	1	1	95	117	154	55	28	128	0	1	1
46	118	168	76	60	188	0	1	0	96	150	172	96	35	235	0	1	0
47	80	152	75	44	173	0	1	1	97	160	160	70	69	170	0	1	0
48	170	160	64	45	143	0	1	1	98	170	160	110	41	174	0	1	1
49	180	160	60	32	128	0	1	1	99	90	154	50	25	175	1	1	1
50	150	162	68	40	342	0	1	1	100	180	152	85	60	280	0	0	1

By using software of Minitab (13.2). We compare between linear Regression model with intercept and without intercept

3.2 The Statistical Analysis

We get the linear regression model by using analysis data. The regression equation is

$$y = 214 - 1.08 X_1 + 0.631 X_2 + 1.39 X_3 + 0.0839 X_4 - 34.1 X_5 + 4.93 X_6 - 11.2 X_7$$

Table (2): Analysis of Variance by using one-way ANOVA table

Source	DF	SS	MS	F	P
Regression	7	76753	10965	8.91	0.000
Residual Error	92	113278	1231		

Total	99	190031			
-------	----	--------	--	--	--

For the purpose of applying the method leverage point. the point (n^*x, n^*y) is added to the original data we know if the new point effect the imported of the intercept in original data as:

$$n^* = 11.049$$

$$n^*\bar{Y} = 1787.94$$

$$n^*\bar{X}_1 = 1772.149$$

$$n^*\bar{X}_2 = 756.414$$

$$n^*\bar{X}_3 = 521.291$$

$$n^*\bar{X}_4 = 2625.684$$

$$n^*\bar{X}_5 = 1.546$$

$$n^*\bar{X}_6 = 8.287$$

$$n^*\bar{X}_7 = 7.403$$

When we added the new compound to original data, the data become $(n=101)$. We get the linear regression model by using analysis data. The regression equation is

$$y = 0.36 + 0.238X_1 + 0.526X_2 + 1.45X_3 + 0.0850X_4 - 27.7X_5 + 2.40X_6 + 1.10X_7$$

Table (3): Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	2687881	383983	297.01	0.000
Residual Error	93	120235	1293		
Total	100	2808116			

We calculate leverage point by equation (9)

$$h_{n+1} = h_{101} = 0.836$$

We show the leverage of this point is large if we compare with the value of

$$\frac{3p}{n} = 0.237$$

Since the $(h_{n+1} > \frac{3p}{n})$ we conclusion that the leverage of the new point.

Effect and force the level regression through the origin point. It means that the linear regression in original data intercept is significant $(B_0 \neq 0)$.

If we test the significant of intercept (B_0) in original data we get the Results

$$t_0 = 2.38$$

$$t_{(0.025,92)} \approx 2.00$$

Since ($t_0 > t_{(0.025,92)}$) it means that the intercept is very necessary in this model. We show for two ways that the significant intercept is very necessary, if we sure that the new points when added the original data forced the level regression through the original point. We test significant intercept in new data by using (t) test in (11) as:

$$t_0 = 0.07$$

$$t_{(0.025,93)} \approx 2.00$$

Since ($t_0 < t_{(0.025,93)}$) it means that the intercept in this model is insignificant it means that the leverage new point when added the original data forced the level regression through the original point.

3.3 Comparison between Models with and without Intercept

We show the analysis variance in table (2) mean square error (MSE) equal (10965) is less than the MSE in table (3) is (383983). It means that the original data is best than the new data when we added the new point. when we test this models with no intercept, and notes that many of the usual statistics (such as R^2 and the model F) are not comparable between the intercept and without intercept models, because if we delete the intercept in linear regression the value of (R^2) is increase, Explains the value of (F) increase.

Conclusion

When compared between models with and without intercept, we show the model with intercept is basted than the model without intercept when we forecast to increase or decrease the blood pressure. We note in application that the data is control the style to be used for analysis and to acceptable results.

References

- [1] B.R. Kirkwood, *Medical Statistics*, (1988), www.blackwell-science.com.
- [2] G. Casella, Leverage and regression through the origin, *The American Statistician*, 37(2) (1983), 147-152.

- [3] J. Cohen, P. Cohen, S.G. West and L.S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences (3rd Edition)*, Mahwah, NJ: Lawrence Erlbaum Associates, (2003).
- [4] D.R. Helsel and R.M. Hirsch, *Techniques of Water-Resources Investigations of the United States Geological Survey Book 4*, Hydrologic Analysis and Interpretation Reston, VA, USA, (1991).
- [5] D.L. Farnsworth, The effect of a single point on correlation and slope, *Internat. J. Math. & Math. Scio.*, 13(4) (1990), 799-806.
- [6] D.R. Helsel and R.M. Hirsch, Statistical methods in water resources techniques of water resources investigations, *U.S. Geological Survey, Book 4 (Chapter A3)* (2002), 168 pages.
- [7] D.C. Hoaglin and R.E. Welsch, The hat matrix in regression and ANOVA, *The American Statistician*, 32(1) (1978), 17-22.
- [8] J.O. Rawlings, Maneesha and P. Bajpai, Multiple regression analysis using ANCOVA in university model, *International Journal of Applied Physics and Mathematics*, 3(5) (2013), 336-340.
- [9] R.A. Johnson, *Probability and Statistics for Engineers (6th ed.)*, Pearson Education, (2003).
- [10] Gujarati, *Basic Econometrics (4th Edition)*, The McGraw–Hill Companies, (2004).