Restor Dent Endod. 2018 Aug;43(3):e34
https://doi.org/10.5395/rde.2018.43.e34
pISSN 2234-7658·eISSN 2234-7666

Restorative
Dentistry
& Endodontics RDE

Open Lecture on Statistics

Check for updates

Statistical notes for clinical researchers: simple linear regression 2 – evaluation of regression line

Hae-Young Kim *

Department of Health Policy and Management, College of Health Science, and Department of Public Health Science, Graduate School, Korea University, Seoul, Korea

*Correspondence to
Hae-Young Kim, DDS, PhD
Professor, Department of Health Policy and Management, Korea University College of Health Science, and Department of Public Health Science, Korea University Graduate School, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Korea.
E-mail: kimhaey@korea.ac.kr

ORCID iDs
Hae-Young Kim
https://orcid.org/0000-0003-2043-2575

In the previous section, we established a simple linear regression line by finding the slope and intercept using the least square method as: $\hat{Y}$=30.79 + 0.71X. Finding the regression line was a mathematical procedure. After that we need to evaluate the usefulness or effectiveness of the regression line, whether the regression model helps explain the variability of the dependent variable. Also, statistical inference of the regression line is required to make a conclusion at the population level, because practically, we work with a sample, which is a small part of population. Basic assumption of sampling method is simple random sampling.

USEFULLNESS OF REGRESSION LINE

Coefficient of determination (R-square)

An important purpose of data analysis is to explain the variability in a dependent variable, Y. To compare models, we consider a basic situation where we have information on the dependent variable only. In this situation, we need to model the mean of Y, $\bar{Y}$, to explain the variability, which is regarded as a basic or baseline model. However, when we have other information related to the dependent variable, such as a continuous independent variable, X, we can try to reduce the (unexplained) variability of dependent variable by adding X to the model and we model a linear regression line formed by the predicted value by X, $\hat{Y}$. Hopefully, we can expect an effective reduction of variability of Y as the result of adding X into the model. Therefore, there is a need for evaluation of the regression line, such as "how much better it is than the baseline model using only the mean of the dependent variable".

Figure 1 depicts baseline and regression models and their difference in variability. The baseline model (**Figure 1A**) explains Y using only the mean of Y, which makes a flat line ($\bar{Y}$). The variability in the baseline model is the residual ① $Y - \bar{Y}$. In **Figure 1B,** X, which has a linear relationship with Y, is introduced as an explanatory variable and a linear regression line is fitted. In the regression model, residual ② $Y - \hat{Y}$ is reduced compared to ①. The explained variability ③ $\hat{Y} - \bar{Y}$ represents the amount of reduced variability of Y due to adopting regression line instead of mean Y.

Table 1 shows the same three quantities ① $Y - \bar{Y}$, ② $Y - \hat{Y}$, and ③ $\hat{Y} - \bar{Y}$, which appear in **Figure 1.** $\bar{Y}$ represents mean of Y and $\hat{Y}$, the predicted value of Y by the regression equation $\hat{Y}$ = 30.79 + 0.71X, which are points on the regression line. To quantify the effect of the regression model, we may use a form of ratio of ① $Y - \bar{Y}$ and ③ $\hat{Y} - \bar{Y}$. To facilitate the

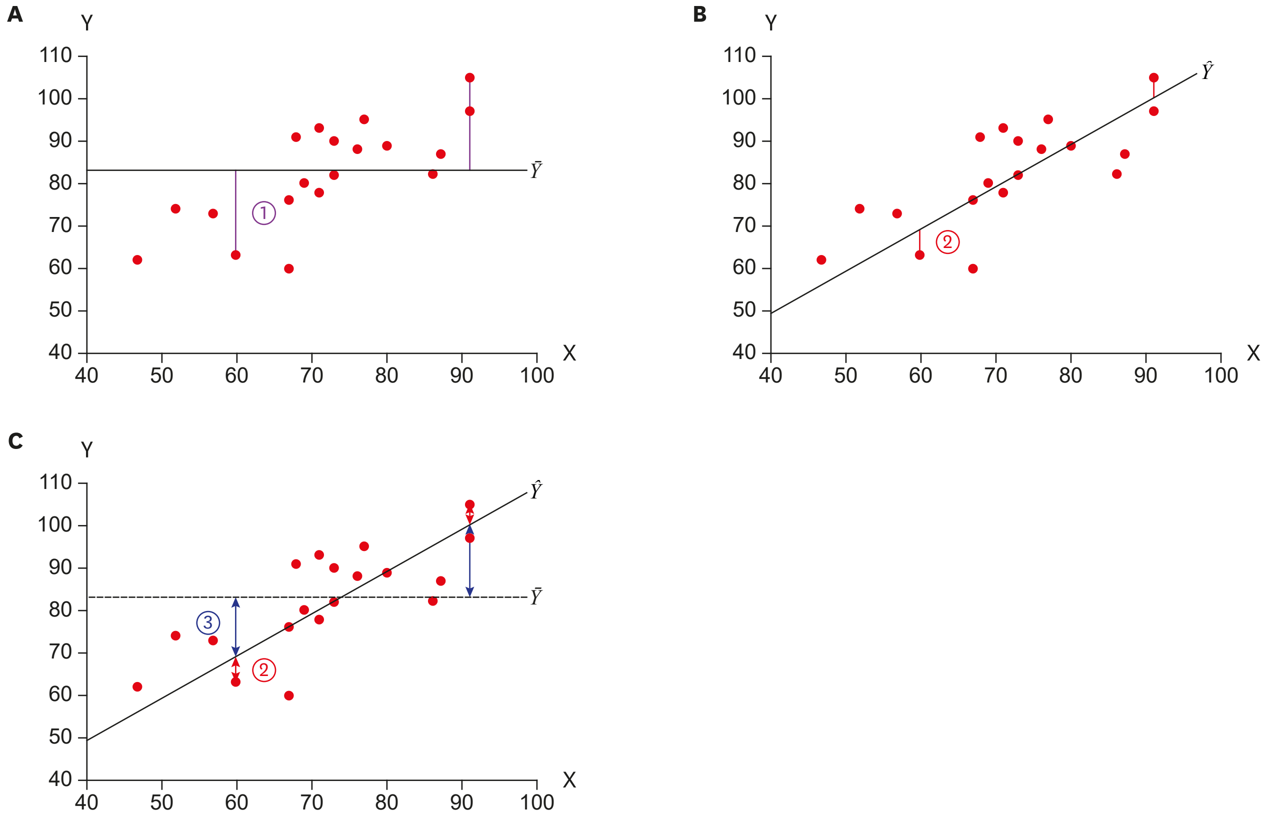

Figure 1. Depiction of comparative models: (A) a baseline model with mean of Y and its residuals (① $Y - \bar{Y}$, total variability); (B) fitted regression line using X and residuals (② $Y - \hat{Y}$, remaining residual); (C) the residual ① in (A) was divided into the explained variability by the line (③ $\hat{Y} - \bar{Y}$, explained variability due to regression) and reduced but remaining unexplained residual (②).

Table 1. Calculation of sum of squares of total (SST), sum of squares due to regression (SSR), sum of squares of errors (SSE), and R-square, which is the proportion of explained variability (SSR) among total variability (SST)

No	X	Y	① $Y - \bar{Y}$	$(Y - \bar{Y})^2$	$\hat{Y}$	③ $\hat{Y} - \bar{Y}$	$(\hat{Y} - \bar{Y})^2$	② $Y - \hat{Y}$	$(Y - \hat{Y})^2$
1	73	90	7.65	58.52	82.74	0.39	0.15	7.26	52.68
2	52	74	−8.35	69.72	67.80	−14.55	211.75	6.20	38.46
3	68	91	8.65	74.82	79.18	−3.17	10.02	11.82	139.62
4	47	62	−20.35	414.12	64.24	−18.11	327.96	−2.24	5.02
5	60	63	−19.35	374.42	73.49	−8.86	78.48	−10.49	110.06
6	71	78	−4.35	18.92	81.32	−1.03	1.06	−3.32	11.01
7	67	60	−22.35	499.52	78.47	−3.88	15.04	−18.47	341.22
8	80	89	6.65	44.22	87.72	5.37	28.87	1.28	1.63
9	86	82	−0.35	0.12	91.99	9.64	92.98	−9.99	99.85
10	91	105	22.65	513.02	95.55	13.20	174.26	9.45	89.29
11	67	76	−6.35	40.32	78.47	−3.88	15.04	−2.47	6.11
12	73	82	−0.35	0.12	82.74	0.39	0.15	−0.74	0.55
13	71	93	10.65	113.42	81.32	−1.03	1.06	11.68	136.46
14	57	73	−9.35	87.42	71.36	−10.99	120.86	1.64	2.70
15	86	82	−0.35	0.12	91.99	9.64	92.98	−9.99	99.85
16	76	88	5.65	31.92	84.88	2.53	6.38	3.12	9.76
17	91	97	14.65	214.62	95.55	13.20	174.26	1.45	2.10
18	69	80	−2.35	5.52	79.90	−2.45	6.03	0.10	0.01
19	87	87	4.65	21.62	92.70	10.35	107.21	−5.70	32.54
20	77	95	12.65	160.02	85.59	3.24	10.49	9.41	88.58
		$\bar{Y} = 82.35$	SST = $\Sigma[Y - \bar{Y}]^2$ = 2,742.55			SSR = $\Sigma[\hat{Y} - \bar{Y}]^2$ = 1,475.04		SSE = $\Sigma[Y - \hat{Y}]^2$ = 1,267.51	
			R^2 = SSR/SST = 1,475.04/2,742.55 = 0.538						

$\hat{Y} = 30.79 + 0.71X$.

calculation procedure, we practically use squared sum, such as $\sum(Y-\bar{Y})^2$ and $\sum(\hat{Y}-\bar{Y})^2$. For the baseline model, we square all the residuals ① $Y - \bar{Y}$ and sum them up, which is called sum of squares of total (SST), $\sum(Y-\bar{Y})^2$. SST is the maximum sum of squares of errors for the data because the minimum information of Y itself was only used for the baseline model. For the regression model, we square all the differences ③ $\hat{Y} - \bar{Y}$ and sum them up, which is called sum of squares due to regression (SSR), $\sum(\hat{Y}-\bar{Y})^2$. SSR is the additional amount of explained variability in Y due to the regression model compared to the baseline model. The difference between SST and SSR is remaining unexplained variability of Y after adopting the regression model, which is called as sum of squares of errors (SSE). SSE can be directly obtained by sum of squares of residual, $\sum(Y-\hat{Y})^2$.

$$\sum(Y-\bar{Y})^2 = \sum(\hat{Y}-\bar{Y})^2 + \sum(Y-\hat{Y})^2$$

$$\text{SST} \quad = \quad \text{SSR} \quad + \quad \text{SSE}$$

If the regression line is flat as in **Figure 1A**, there is no contribution of X. Therefore, regression line will be of no use in explaining Y in this case. Larger absolute values of $\hat{Y} - \bar{Y}$ mean larger contribution of the regression line. To quantify the contribution of the regression line, we use ratio of SSR and SST. We call the ratio as R-square, which is also called 'coefficient of determination'. The coefficient of determination is interpreted as the proportion of additional explanation of variability in the regression model among the total variability. The proportion, R-square, is also frequently expressed as a percentage. As below, we can conclude 53.8% of SST is explained by the estimated regression line to predict the dependent variable.

$$\text{Coefficient of determination or R} - \text{square } (R^2) = \frac{SSR}{SST} = \frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2} = \frac{1{,}475.04}{2{,}742.55} = 0.538 \text{ or } 53.8\%.$$

We use the term 'fit' which indicates whether or not the model fits well with the observed data. In this case, the regression model using X fits well because it explains a big amount of variability among SST. A large R-square value represents a good fit.

Overall F test: a global evaluation of model

As mentioned above, SST is divided into SSR and SSE. A relatively small SSE can be interpreted as a "good fit" of the model. The usefulness of the regression model is tested using F test as a global evaluation of the regression model. In the F test, F value is defined as the ratio of mean of squares of regression (MSR) and mean of squares of error (MSE), which are under the title of 'Mean squares' in **Table 2**. Mean squares are defined as the means of SSR or SSE per one degree of freedom (df) and are obtained by dividing SSR or SSE by their own df as shown in **Table 2**. A large F value represents relatively large MSR and relatively small MSE, which can be interpreted as a good contribution of the regression model. A large F value leads to the rejection of null hypothesis that all the slopes are zero. If all the slopes are zero, the regression line is useless because it is identical to the simple mean of Y, such that $\hat{Y} = \bar{Y}$.

The formal expression of the overall test is as follows:

$$H_0\!:\beta_1 = \beta_2 = \cdots = \beta_k = 0 \text{ vs } H_a\!: \textit{at least one } \beta_j \neq 0.$$

Test statistic: $F = \text{MSR/MSE}$

Table 2. A general form of analysis of variance table for the global evaluation of the regression line

Source	Sum of squares	df	Mean squares	F
Regression	SSR = $\Sigma(\hat{Y} - \bar{Y})^2$	p	MSR = SSR/p	MSR/MSE
Residual	SSE = $\Sigma(Y - \hat{Y})^2$	n – p – 1	MSE = SSE/(n – p – 1)	
Total	SST = $\Sigma(Y - \bar{Y})^2$	n – 1		

df, degrees of freedom; SSR, sum of squares due to regression; MSR, mean of squares of regression; MSE, mean of squares of error; SSE, sum of squares of error; n, number of observation; p, number of predictor variables (Xs) in the model; SST, sum of squares of total.

Table 3. An analysis of variance (ANOVA) table for the example data

		ANOVA			
Model	Sum of squares	df	Mean square	F	*p* value
Regression	1,475.036	1	1,475.036	20.947	<0.001
Residual	1,267.514	18	70.417		
Total	2,742.550	19			

Table was cited from appendix of reference [1] which has modified as table format.
df, degree of freedom.

Table 3 shows the analysis of variance table for the example data. The F value is calculated as follows:

$$F = \text{MSR}/\text{MSE} = 1{,}475.036/70.147 = 20.947$$

To get *p* values corresponding to the calculated F values, we can refer to a website named '*p*-value from F-ratio Calculator' at http://www.socscistatistics.com/pvalues/fdistribution.aspx. We provide 'F-ratio value' as 20.947, 'df-numerator (df of MSR)' as 1, and 'df-denominator (df of MSE)' as 18 in **Table 3**. Then finally, we get the *p* value as 0.000234, which can be expressed as $p < 0.001$, as well. We reject null hypothesis that the regression model is useless and conclude that there is at least one non-zero slope, which means the regression model is contributing in reduction of error.

EVALUATION OF INDIVIDUAL SLOPE(S)

T test: test on significance of slope

If the global evaluation of the regression model concludes there is at least one nonzero slope, we want to know which slope is nonzero and its estimated size. In simple linear regression with only one X, the result of global F test and the significance of the slope share the same conclusion. When we have a simple linear regression model such as $Y = b_0 + b_1X + e$ (residual), the formal test on the significance of slope is as follows:

$$H_0{:}\beta_1 = 0 \text{ vs } H_a{:}\beta_1 \neq 0.$$

Test statistic: $t = \dfrac{b_1}{se_{b_1}}$, where se_{b_1} is the standard error of b_1.

This test is asking whether a statistically significant linear relationship exists between the dependent and independent variables. If the answer is no, the regression model is no good because the basic assumption was a linear relationship between two variables. If there is no linear relationship, the slope is near zero and the scatterplot of Y and X is expected to show a random scatter, which means there is no correlation.

Table 4 displays information on regression coefficients from the example data from the previous section [1]. Information on slope is in the line of X. The point estimate b_1 is 0.712 and its standard error (se_{b_1}) is 0.155. The standard error of b_1 (se_{b_1}) and test statistic are obtained as follows.

Table 4. Regression coefficients from the example data

Coefficients							
Model	Unstandardized coefficients		Standardized coefficients	*t*	*p* value	95% CI for bound	
	B	SE	Beta			Lower bound	Upper bound
1 (Constant)	30.795	11.420		2.697	0.015	6.803	54.787
X	0.712	0.155	0.733	4.577	0.000	0.385	1.038

Table was cited from appendix of reference [1] which has modified as table format.
SE, standard error; CI, confidence interval.

$$se_{b_1} = \frac{\sqrt{MSE}}{\sqrt{\sum(x_i-\bar{x})^2}} = \frac{\sqrt{70.417}}{\sqrt{2,912.95}} = 0.155,$$

where $\sum(x_i - \bar{x})^2 = 2,912.95$ from **Table 1** in the previous section [1].

$$t = \frac{b_1}{s_{b_1}} = \frac{0.712}{0.155} = 4.577 \sim t_{n-2=18} \qquad p < 0.001$$

To get *p* values corresponding to the calculated *t* values, we can refer to a website named '*p* value from T score calculator' at http://www.socscistatistics.com/pvalues/tdistribution.aspx. We provide 'T score' as 4.577, and 'DF' as 18, which is df of MSE, n-2, and then finally we get the *p* value as 0.000234, which is the same *p* value from F test above. We reject null hypothesis that the slope is zero and conclude that there is a significant linear relationship between Y and X.

95% confidence interval (CI) of slope

As a statistical inference procedure, we usually use an interval estimation procedure. We got the estimate using a sample. But if we take different sample from the population, we may have different estimate. The phenomenon is called 'sampling variability'. Therefore, we usually make CI in which we believe the parameter β_1 is included with certain degree of confidence.

The 95% CI of β_1 is expressed as $b_1 \pm t_{df=18,0.025} {}^* se_{b_1}$. We already know b_1 and se_{b_1}, shown in **Table 4**. We can get the critical *t* value with df = 18 from a website 'Free critical *t* value calculator' at https://www.danielsoper.com/statcalc/calculator.aspx?id=10. We provide 'Degrees of freedom' as 18, 'Probability level' as 0.05, and get *t* value (2-tailed) as 2.100. We are ready to construct the 95% CI of β_1 as follows:

$$b_1 \pm t_{df=18,0.025} {}^* se_{b_1},$$

$$0.712 \pm 2.100 * 0.155,$$

$$0.712 \pm 0.326,$$

$$(0.386, 1.038)$$

Does the interval contain zero? No! Therefore, we can conclude that the slope of X is nonzero similar to the result from the *t* test above.

REFERENCES

1. Kim HY. Statistical notes for clinical researchers: simple linear regression 1 – basic concepts. Restor Dent Endod 2018;43:e21.
PUBMED | CROSSREF