

SOME COMMENTS ON REGRESSION WITHOUT INTERCEPT

In Problem 13 of Chapter 2, we are interested in comparing simple linear regression analysis with or without intercept. Below are the results from SAS using the dataset `sriver`.

Why are R^2 and F statistics for the no-intercept fit much higher than those with the intercept? It is even more peculiar as the Error SS of the no-intercept fit is larger than that with the intercept. Let us explain this by answering a series of questions below.

1. For the fit with intercept, what are the Total SS (SST) and the Model SS (SSR)? How are they computed?
2. For the no-intercept fit, what are the SST and the SSR? How are they computed?

The REG Procedure
 Model: MODEL1
 Dependent Variable: WaYield

Number of Observations Read	17
Number of Observations Used	17

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	307.25750	307.25750	101.16	<.0001
Error	15	45.56015	3.03734		
Corrected Total	16	352.81765			

Root MSE	1.74280	R-Square	0.8709
Dependent Mean	15.71176	Adj R-Sq	0.8623
Coeff Var	11.09231		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.72538	1.54882	0.47	0.6463
WaCont	1	0.49808	0.04952	10.06	<.0001

The REG Procedure
Model: MODEL2
Dependent Variable: WaYield

Number of Observations Read 17
Number of Observations Used 17

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4503.20362	4503.20362	1558.66	<.0001
Error	16	46.22638	2.88915		
Uncorrected Total	17	4549.43000			

Root MSE 1.69975 R-Square 0.9898
Dependent Mean 15.71176 Adj R-Sq 0.9892
Coeff Var 10.81832

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
WaCont	1	0.52039	0.01318	39.48	<.0001

1. For the fit with the intercept, SST is the corrected SS for the response, which can be interpreted as the Error SS (SSE) for the model $y = \mu + \epsilon$. The SSR shows the reduction from SST due to fitting the regression model. In this example,

$$\text{SST} = 352.81765, \quad \text{SSR} = 307.25750, \quad R^2 = \text{SSR} / \text{SST} = 0.8709.$$

2. However, in the no-intercept fit, SST is the uncorrected SS, which can be interpreted as the Error SS for fitting the model $y = 0 + \epsilon$, a very poor model! The Model SS is the reduction from that sum due to fitting the no-intercept regression model. Thus

$$\text{SST} = 4549.43000, \quad \text{SSR} = 4503.20362, \quad R^2 = \text{SSR} / \text{SST} = 0.9898,$$

which is larger than that for the intercept fit because the reduction starts from a much larger value.

Note the differences in the SAS output for the no-intercept fit, especially the line at the top:

NOTE: No intercept in model. R-Square is redefined.

Moreover, the no-intercept line does not go through the means of the variables. Remember the test statistic for the model is based on the difference between the SSE and SST. The former measures the variability from the estimated regression, while the latter measures the variability from the null model.

In the model with intercept, the null hypothesis $\beta_1 = 0$ specifies the model $y = \beta_0 + \epsilon = \mu + \epsilon$. Now μ is estimated by the sample mean of the responses. Therefore the test for the model compares the variability from the regression to the variation from the sample mean.

$$\begin{aligned} \text{SST} &= \text{SSE} + (\text{SST} - \text{SSE}), \\ \sum_{i=1}^n (Y_i - \bar{Y})^2 &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2. \end{aligned}$$

In the no-intercept case, the null hypothesis $\beta_1 = 0$ specifies the model $y = \epsilon$. The test compares the variation from the regression to that from $y = 0$. Obviously, unless the mean of the responses is close to zero, the

variation from the mean is smaller than the variation from zero, hence the apparent contradiction.

$$\sum_{i=1}^n Y_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n \hat{Y}_i^2,$$

Uncorrected SS = Error SS + Model SS.