Lecture 20: Regression through the Origin and Its Pitfalls

In this lecture we cover regression through the origin. Sometimes, usually not often, the regression function is linear and goes through the origin. Examples of relationships having zero intercept are age of a plant and its height, length of a leaf and its weight etc.

The linear model with no intercept has the form

$Y_i = \beta_1 X_i + \varepsilon_i$

where β_1 = slope parameter

X_i = known constants

$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

Under the assumptions of a linear regression model, the least squares estimation leads to the minimization, with respect to β_1 of

$Q = \Sigma(Y_i - \beta_1 X_i)^2$

Differentiating this function with respect to β_1 and equating to zero

$$\frac{dQ}{d\beta_1} = \Sigma 2(Y_i - \beta_1 X_i)(-X_i) = 0$$

$$\Sigma\{X_i(Y_i - \beta_1 X_i)\} = 0$$

or $$\Sigma X_i Y_i - \beta_1 \Sigma X_i^2 = 0$$

and $$b_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2}$$

The regression model is $\hat{Y} = b_1 X$

Note: This is NOT $$b_1 = \frac{\Sigma n X_i Y_i - \Sigma X_i \Sigma Y_i}{\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{s_{XY}}{s_X^2}$$

b_1 also can be shown to be the maximum likelihood estimator of β_1

Unbiased estimator of $E(Y)$ is $\hat{Y}$

$E(\hat{Y}) = E(b_1 X) = XE(b_1) = \beta_1 X = E(Y)$

and an unbiased estimator of σ^2 is

$$MSE = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-1}$$

Variances, Confidence and Prediction Intervals

The variances, confidence and prediction intervals are computed in a completely analogous manner to the simple linear regression model with intercept. That is, we utilize the s(.) and t values to construct our intervals. The variances are defined below.

Estimator	Variance	Confidence/prediction interval
b_1	$s^2(b_1) = \frac{MSE}{\Sigma X_i^2}$	$b_1 \pm t_{(1-\alpha/2,n-1)}s(b_1)$
$\hat{Y}_h$	$s^2(\hat{Y}_h) = MSE\left[\frac{X_h^2}{\Sigma X_i^2}\right]$	$\hat{Y}_h \pm t_{(1-\alpha/2,n-1)}s(\hat{Y}_h)$
$Y_{h\text{-new}}$	$s^2(Y_{h\text{-new}}) = MSE\left[1+\frac{X_h^2}{\Sigma X_i^2}\right]$	$Y_{h\text{-new}} \pm t_{(1-\alpha/2,n-1)}s(Y_{h\text{-new}})$

The major differences in regression without and with intercept are:

1. The degrees of freedom for residuals are (n-1) instead of (n-2). Why?

2. As the regression is forced through (0,0), $\Sigma(X_i - \overline{X})^2$ is replaced by ΣX_i^2 and $(X_h - \overline{X})^2$ by X_h^2.

3. The coefficient of determination

$$R^2 = \frac{SSR}{SSTOU} = \frac{\Sigma\hat{Y}^2}{\Sigma Y^2} = \frac{b_1^2\Sigma X^2}{\Sigma Y^2}$$

can be absurdly large even when the correlation between X and Y is weak. Being totally meaningless, it is not even mentioned in most computer outputs.

The ANOVA Table for Regression Through the Origin

SOURCE OF VARIATION	SUM OF SQUARE	DEGREES OF FREEDOM	MEAN SQUARE	E(MS)	F
REGRESSION	$SSR = \Sigma \hat{Y}_i^2$	1	MSR = SSR/1	$\sigma^2 + \beta_1^2 \Sigma X_i^2$	MSR/MSE
ERROR	$SSE = \Sigma(Y_i - \hat{Y}_i)^2$	n - 1	MSE = SSE/(n-1)	σ^2	
TOTAL	$SSTOU = \Sigma Y_i^2$	n			

Exercise: The instrument data set provides a good example when one might specify a model without an intercept. Try it!

Example

The Sturgeon research is an example where it makes no sense. Why not? (Because zero fish length observations are not within the scope of the mode). We may still be tempted to do regression through the origin. Let's illustrate why it would be meaningless.

From previous, we have

$\Sigma X = 1{,}502$; $\Sigma X^2 = 234{,}580$; $\Sigma XY = 258{,}031$

$\Sigma Y = 1{,}562$; $\Sigma Y^2 = 305{,}982$; $n = 10$

$$b_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2} = \frac{258{,}031}{234{,}580} = 1.09997$$

Hence $\hat{Y} = b_1 X = 1.09997X$

$$SSE = \Sigma(Y_i - \hat{Y}_i)^2 = \Sigma Y_i^2 - (b_1)^2 \Sigma X_i^2 = 305{,}982 - (1.09997)^2(234580)$$

$$= 22{,}155.68$$

$$\text{MSE} = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-1} = \frac{22155.68}{9} = 2461.74$$

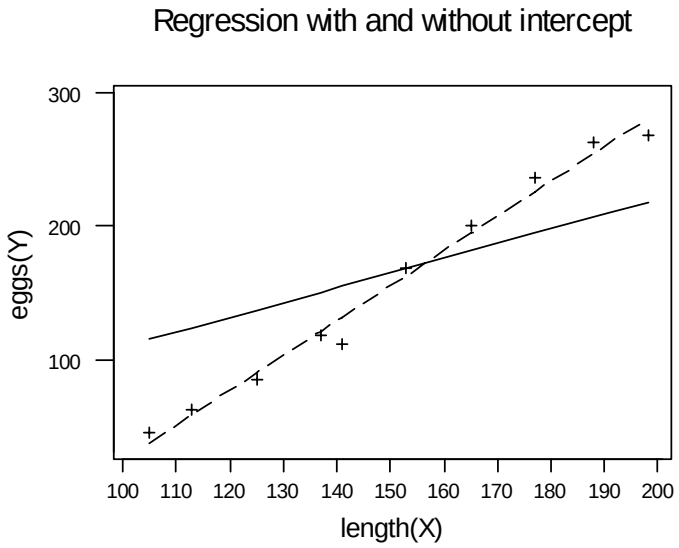
(where before it was equal to 115)!

and $$s^2(b_1) = \frac{\text{MSE}}{\Sigma X_i^2} = \frac{2461.74}{234{,}580} = 0.01049$$

or $$s(b_1) = 0.1024$$

The regression lines, through the origin (solid line) and with the intercept (dashed line), are in Fig 5.2.

Fig 5.2



The 95% CI for β_1 is

$$b_1 \pm t_{(0.975,9)}s(b_1) = 1.09997 \pm 2.262*0.1024$$

or $$0.868 \leq \beta_1 \leq 1.332$$

Note df = 9 or n -1and not 10 or n – 2.

ANOVA table for regression through the origin Sturgeon data

SOURCE OF VARIATION	SUM OF SQUARE	DEGREES OF FREEDOM	MEAN SQUARE	F
REGRESSION	283,826	1	283,826	115.3**
ERROR	22,156	9	2,462	
TOTAL	305,982	10		

As mentioned earlier, based on the above data, the coefficient of determination is

$$R^2 = \frac{SSR}{SSTOU} = \frac{283{,}826}{305{,}982} = 0.928$$ and is not accurate.

It is more meaningful to compute R^2 for regression through the origin, as 1 minus the ratio of SSE from the regression, and SSTO corrected for the mean. That is

$$R^2 = 1 - \frac{\Sigma(Y - b_1X)^2}{\Sigma(Y - \overline{Y})^2}$$

$$= 1 - \frac{22{,}156}{61{,}997.6}$$

$$= 0.6426$$

The 95% confidence boundaries for $E\{Y_h\}$ has a different look

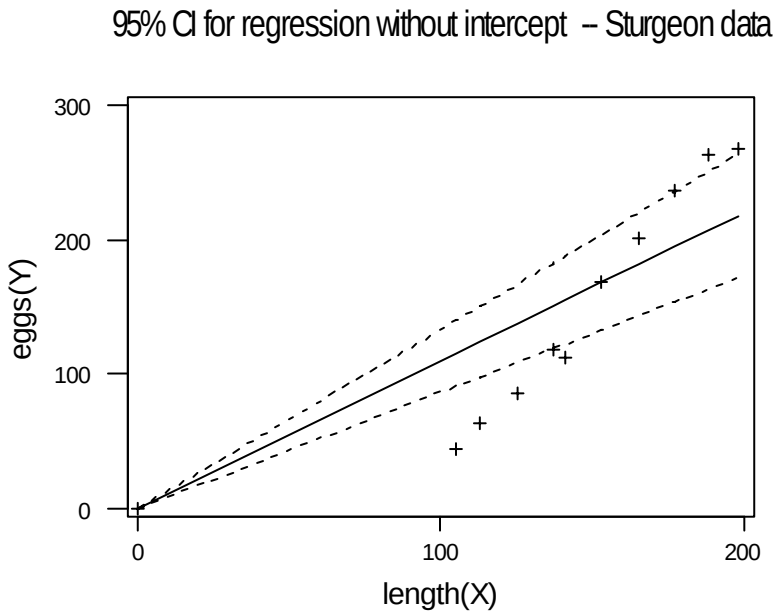


Fig 5.3

Notes:

Regression through the origin should not be forced unless there are very compelling reason to do so.

If the data has not been collected in the vicinity of X = 0, forcing through the origin, is likely to make the regression worse as can be seen in the Sturgeon example.

The sturgeon data with residual (e_i) and $\hat{Y}_i$

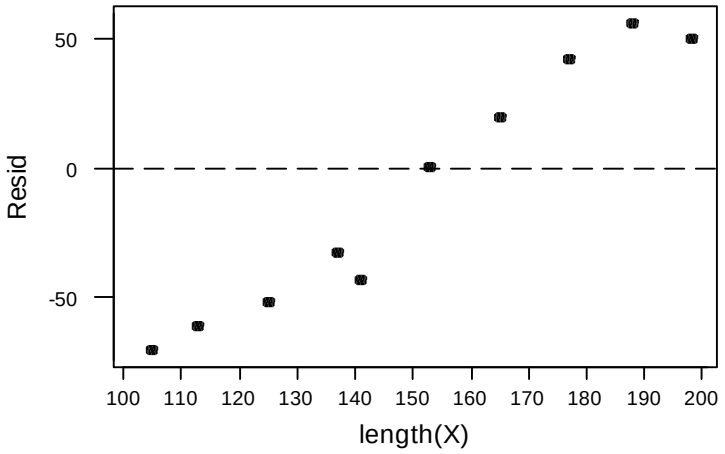
length(X_i)	eggs(Y_i)	e_i	$\hat{Y}_i$
105	45	-70.50	115.50
113	63	-61.30	124.30
125	86	-51.50	137.50
137	118	-32.70	150.70
141	112	-43.10	155.10
153	169	0.70	168.30
165	201	19.50	181.50
177	237	42.30	194.70
188	263	56.21	206.79
198	268	50.21	217.79

Notes:

Unlike in regression with intercept $\Sigma e_i \neq 0$, but $\Sigma e_i X_i = 0$

X_i (or $\hat{Y}_i$) and e_i are correlated (see Fig 5.4)

Residual plot - Sturgeon data,,regression without intercept



Forcing the regression through any other point {except ($\overline{X}, \overline{Y}$)} will always make the MSE larger. Why? The line will be off center.

Example Instrument Data

Variables Entered/Removed[b,c]

Model	Variables Entered	Variables Removed	Method
1	instrument 2[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: instrument 1

c. Linear Regression through the Origin

Model Summary

Model	R	R Square[a]	Adjusted R Square	Std. Error of the Estimate
1	.999[b]	.998	.998	1.9604

a. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

b. Predictors: instrument 2

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12573.099	1	12573.099	3271.634	.000[a]
	Residual	26.901	7	3.843		
	Total	12600.000[b]	8			

a. Predictors: instrument 2

b. This total sum of squares is not corrected for the constant because the constant is zero for regression through the origin.

c. Dependent Variable: instrument 1

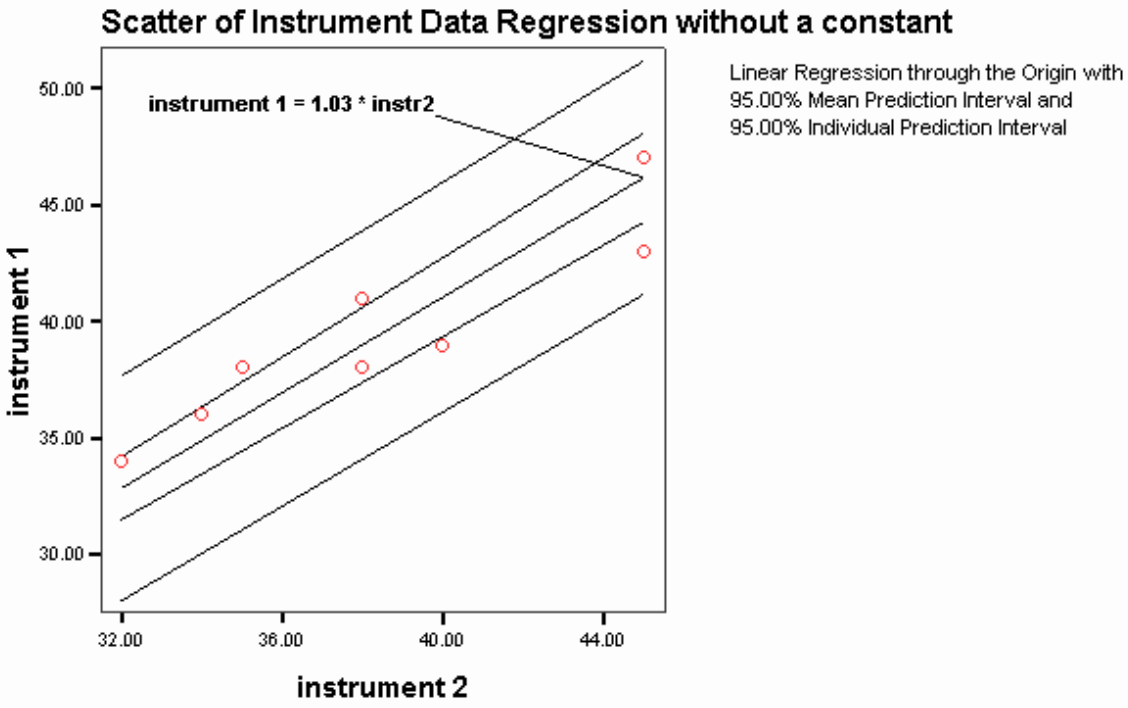
d. Linear Regression through the Origin

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	instrument 2	1.026	.018	.999	57.198	.000	.984	1.068

a. Dependent Variable: instrument 1

b. Linear Regression through the Origin


Scatter of Instrument Data Regression without a constant
instrument 1 = 1.03 * instr2
Linear Regression through the Origin with
95.00% Mean Prediction Interval and
95.00% Individual Prediction Interval
50.00
45.00
40.00
35.00
30.00
instrument 1
32.00
36.00
40.00
44.00
instrument 2