Linear Regression Evaluation of Model Estimators

Various Metrics

Metrics

- MSE
- Karl Pearson's Coefficients of Correlation r
- Computation of R^2
- Multiple R
- Standard Error of Estimate

Calculating Residual and MSE for Regression

For the small data set calculate the residuals and the estimate for σ^2 given the LSRL:

$\hat{y} = 1.5 + 1.5x$

Residual: $y - \hat{y}$

x	y	$\hat{y}$
1	3	3
2	4	4.5
2	5	4.5
3	6	6

x=2

Y=3

$\hat{y} = 1.5 + 1.5(1)$ x=1

$\hat{y} = 1.5 + 1.5 = 3$

$\hat{y} = 1.5 + 1.5(2)$

$\hat{y} = 1.5 + 3$

$\hat{y} = 4.5$

$\hat{y} = 1.5 + 1.5(3)$

$= 1.5 + 4.5$

$= 6$

For the small data set calculate the residuals and the estimate for σ^2 given the LSRL:

$\hat{y} = 1.5 + 1.5x$

Residual: $y - \hat{y}$

x	y	$\hat{y}$	Residual $y - \hat{y}$
1	3	3	$3 - 3 = 0$
2	4	4.5	$4 - 4.5 = -0.5$
2	5	4.5	$5 - 4.5 = 0.5$
3	6	6	$6 - 6 = 0$

$x = 2$

$Y = 3$

$\hat{y} = 1.5 + 1.5(1)$ $x = 1$

$\hat{y} = 1.5 + 1.5 = 3$

$\hat{y} = 1.5 + 1.5(2)$

$\hat{y} = 1.5 + 3$

$\hat{y} = 4.5$

$\hat{y} = 1.5 + 1.5(3)$

$= 1.5 + 4.5$

$= 6$

For the small data set calculate the residuals and the estimate for σ^2 given the LSRL:

$\hat{y} = 1.5 + 1.5x$

Residual: $y - \hat{y}$

x	y	$\hat{y}$	Residual $y-\hat{y}$
1	3	3	$3-3=0$
2	4	4.5	$4-4.5=-0.5$
2	5	4.5	$5-4.5=0.5$
3	6	6	$6-6=0$

$x=2$

$x=3$

$$\hat{y} = 1.5 + 1.5(1) \quad x=1$$
$$\hat{y} = 1.5 + 1.5 = 3$$

$$\hat{y} = 1.5 + 1.5(2)$$
$$\hat{y} = 1.5 + 3$$
$$\hat{y} = 4.5$$

$$\hat{y} = 1.5 + 1.5(3)$$
$$= 1.5 + 4.5$$
$$= 6$$

$$MSE = s^2 = \frac{\sum (y-\hat{y})^2}{n-(k+1)}$$
$$= \frac{\sum (\text{residual})^2}{n-(k+1)}$$

However, a terminological difference arises in the expression mean squared error (MSE). The mean squared error of a regression is a number computed from the sum of squares of the computed residuals, and not of the unobservable errors. If that sum of squares is divided by n, the number of observations, the result is the mean of the squared residuals. Since this is a biased estimate of the variance of the unobserved errors, the bias is removed by dividing the sum of the squared residuals by $df = n - p - 1$, instead of n, where df is the number of degrees of freedom (n minus the number of parameters (excluding the intercept) p being estimated - 1). This forms an unbiased estimate of the variance of the unobserved errors, and is called the mean squared error.[4]

[4] Steel, Robert G. D.; Torrie, James H. (1960). Principles and Procedures of Statistics, with Special Reference to Biological Sciences. McGraw-Hill. p. 288.

Ref: https://en.wikipedia.org/wiki/Errors_and_residuals

For the small data set calculate the residuals and the estimate for σ^2 given the LSRL:

$\hat{y} = 1.5 + 1.5x$

Residual: $y - \hat{y}$

	x	y	$\hat{y}$	Residual $y-\hat{y}$
	1	3	3	$3-3=0$
$x=2$	2	4	4.5	$4-4.5=-0.5$
	2	5	4.5	$5-4.5=0.5$
$y=3$	3	6	6	$6-6=0$

$$MSE = s^2 = \frac{\sum(y-\hat{y})^2}{n-(k+1)} = \frac{\sum(\text{residual})^2}{n-(k+1)}$$

$$= \frac{0^2 + (-0.5)^2 + (0.5)^2 + 0^2}{2}$$

$$= \frac{0 + 0.25 + 0.25 + 0}{2}$$

$$= \frac{0.5}{2} = \boxed{0.25}$$

Karl Pearson's Coefficients of Correlation r

For this purpose, ...

7.4.1 Karl Pearson's Coefficient of Correlation

Karl Pearson's coefficient of correlation is a helpful statistical formula that quantifies the strength between two variables. This coefficient value helps in determining how strong that relationship is between the two variables. The Pearson coefficient is given by

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}} \tag{7.14}$$

where x and y are variables and N is the number of instances we have to compute the coefficient.

It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables.

Table 7.2 Dataset of height and weight observations

S. No.	Height (X) cm	Weight (Y) kg	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$
1	151	63	−2.8	−2.3	6.44	7.84
2	174	81	20.2	15.7	317.14	408.04
3	138	56	−15.8	−9.3	146.94	249.64
4	186	91	32.2	25.7	827.54	1036.8
5	128	47	−25.8	−18.3	472.14	665.64
6	136	57	−17.8	−8.3	147.74	316.84
7	179	76	25.2	10.7	269.64	635.04
8	163	72	9.2	6.7	61.64	84.64
9	152	62	−1.8	−3.3	5.94	3.24
10	131	48	−22.8	−17.3	394.44	519.84
	$\bar{X} =$ 153.8	$\bar{Y} =$ 65.3			$\Sigma = 2649.6$	$\Sigma = 3927.6$

Solution:

Now

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{2649.6}{3927.6} = 0.6746$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X} = 65.3 - (0.6746 \times 153.8) = -38.4551$$

S. No.	Height (x)	Weight (y)	xy	x^2	y^2
8	163	72	11736	26569	5184
9	152	62	9424	23104	3844
10	131	48	6288	17161	2304
			Σ	Σ	Σ
	$\bar{x} = 153.8$	$\bar{y} = 65.3$	103081	240472	44513

Solution:

We will evaluate the Karl Pearson's coefficient to find out if there is a strong relationship between the height and weight of a person. Thus, we can consider the linear regression equation to evaluate weight of the person from the height.

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$r = \frac{(10\times 103081) - (1538\times 653)}{\sqrt{\left[(10\times 240272) - 1538^2\right]\left[(10\times 44513) - 653^2\right]}}$$

$$r = 0.9771$$

The main advantage of this coefficient is that it summarizes in one value the degree and direction of correlation. The limitations of the Pearson's coefficient are listed below:

1. It always assumes linear relationship.
2. Interpreting the value of r is difficult.
3. Value of the correlation coefficient is affected by extreme values.
4. It is time consuming.

R^2

7.4.2 R-Square

R-square gives information about the goodness-of-fit measure for linear regression models. It indicates percentage of variance in the dependent–independent variable pair. It measures the strength of the relationship in a 0 to 100% scale. For each observation in our data, we can compute

$$\hat{y}_i = \bar{y}_i = b_0 + b_1 x_i$$

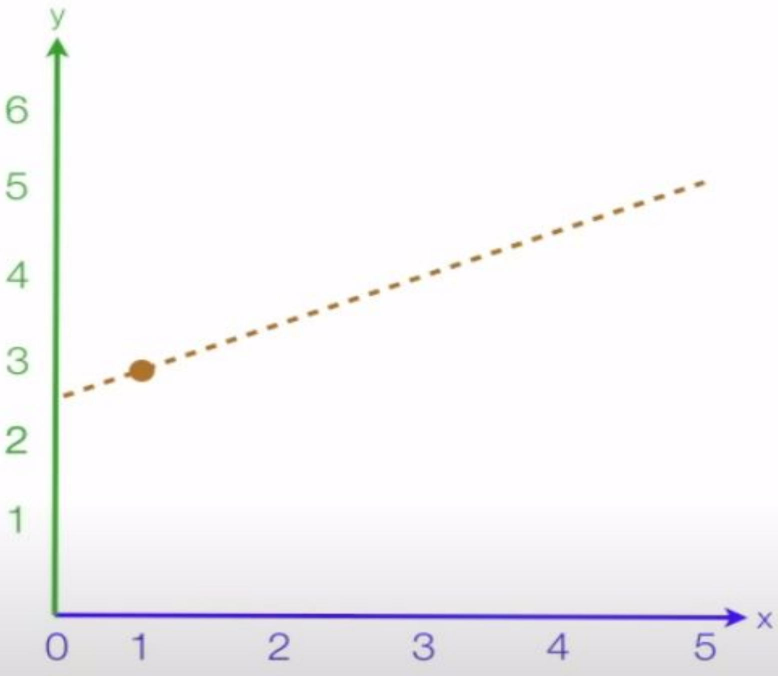
These are called *fitted values*. Thus, the i^{th} fitted value, $\bar{y}_i$, is the point on the least squares regression line corresponding to x_i. For the i^{th} observation, we can compute ordinary least squares residuals as shown in Eq. (7.15).

$$e_i = y_i - \bar{y}_i = y_i - \hat{y}_i \tag{7.15}$$

One of the properties of the residuals is that their sum is zero.

Bo=2.2, B1=0.6, Y bar =4

Row 1= 2 - 4 = -2
when x=1,   2.2 + .6 *1 = 2.8
Row 2 = 4 - 4 = 0
when x=2,   2.2 + .6 *2 = 3.4

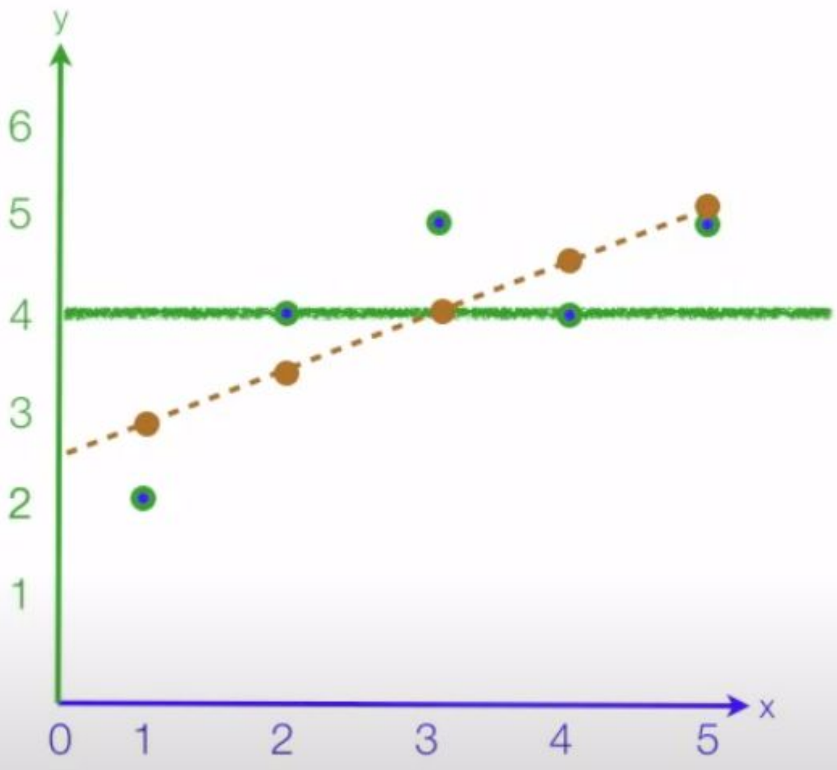


x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$
1	2	-2	4	2.8
2	4	0	0	
3	5	1	1	
4	4	0	0	
5	5	1	1	
			6	
mean	4			

$\hat{y} = 2.2 + .6 \bullet 1$
$= 2.2 + .6$

Ref: https://youtu.be/w2FKXOa0HGA

Bo=2.2, B1=0.6
2.2 + .6 *2 = 3.4



x	y	$y - \bar{y}$	$(y - \bar{y})^2$	$\hat{y}$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	2	-2	4	2.8	-1.2	1.44
2	4	0	0	3.4	-.6	.36
3	5	1	1	4	0	0
4	4	0	0	4.6	.6	.36
5	5	1	1	5.2	1.2	1.44
			6			3.6
mean	4					

$$R^2 = \frac{3.6}{6} = .6 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

The following quantities are also computed.

$$SST = \sum (y_i - \bar{y})^2$$
$$\boxed{SSR = \sum (\hat{y}_i - \bar{y})^2}$$
$$SSE = \sum (y_i - \hat{y}_i)^2 \tag{7.16}$$

where SST is the total sum of squared deviations in y from its mean. SSR is the sum of squares due to regression. SSE is the sum of squared residuals (errors).

The ratio $R^2 = SSR/SST$ can be interpreted as the proportion of the total variation in y that is accounted for by the predictor variable x. The high value of R^2 indicates a strong linear relationship.

Let us consider the following example of dataset of observations (Table 6.4) to compute R^2.

$\bar{y} = 65.3$ = mean of y values.

Table 7.4 The observations of height and weight to compute R^2

Height (cm)	*Weight (kg)*	$\bar{y} = \hat{y}$	*SSR*	*SST*	*SSE*
151	63	63.411	3.56828322	5.29	0.16892922
174	81	78.927	185.696219	246.49	4.29716316
138	56	54.641	113.612576	86.49	1.84666357
186	91	87.022	471.860924	660.49	15.82162
128	47	47.895	302.934721	334.89	0.8009892
136	57	53.292	144.195426	68.89	13.7503023
179	76	82.3	289.00306	114.49	39.691134
163	72	71.506	38.5185321	44.89	0.24371007
152	62	64.086	1.47471878	10.89	4.34981078
131	48	49.919	236.581006	299.29	3.68183182
			1787.44547	1872.1	84.6521541

$$R^2 = SSR/SST = 1784.45/1872.1 = 0.9548$$

Calculations SSR, SST and SSE

Linear Regression Evaluation Metrics

$\hat{Y}_i = b_0 + b_1 x_i$ } These are called fitted values.

$\hat{Y}_1 = b_0 + b_1 x_1$

computed $b_0 = -38.4551$

$b_1 = 0.6746$

$\boxed{x_1 = 151}$

0.6746×151

$\hat{Y}_1 = -38.4551 + \cancel{114.682}$

$\boxed{\hat{Y}_1 = \cancel{76.23} \quad 63.4095 = 63.41}$

for R^2 computation

	Ht(cm) x	wt(kg) y	$\bar{Y}$	SSR	SST	SSE
1.	151	63	63.41	3.56828322	5.29	0.1689

$$SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$$

$$= (\hat{Y}_1 - \bar{Y})^2 = (63.41 - 65.3)^2 = 3.5682$$

$$(63.411 - 65.3)^2 = 3.5682$$

$\bar{Y} = Y$ mean $= 65.3$

$$SST = \Sigma(Y_i - \bar{Y})^2$$

$$= (63 - 65.3)^2 = (-2.3)^2 = 5.29$$

$$SSE = \Sigma(Y_i - \hat{Y}_i)^2 = (63 - 63.41)^2 = (0.41 \times 0.41)^2 = 0.1681$$

Multiple R

131	48	49.919	236.581006	299.29	3.68183182
			1787.44547	1872.1	84.6521541

$$R^2 = SSR/SST = 1784.45/1872.1 = 0.9548$$

7.4.2.1 Multiple R

Multiple R is a correlation coefficient. It gives us an idea of the strength of a linear relationship. For example, a value of 1 means a perfect positive relationship and 0 means no relationship at all. It is the square root of R squared.

$$\text{Multiple R} = \sqrt{R^2}$$

In the above example,

$$\text{Multiple R} = \sqrt{R^2} = 0.9771$$

7.4.3 Standard Error of Estimate

The standard error of the estimate is a measure of the accuracy of predictions. It is given by

$$\sigma_{\text{est}} = \sqrt{\frac{\sum (Y - Y')^2}{N}} \tag{7.17}$$

Standard Error of Estimate

7.4.3 Standard Error of Estimate

The standard error of the estimate is a measure of the accuracy of predictions. It is given by

$$\sigma_{\text{est}} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

• 133

The denominator is the sample size reduced by the number of model parameters estimated from the same data, $(n-p)$ for p regressors or $(n-p-1)$ if an intercept is used. In this case, $p=1$ so the denominator $n-2$.

In the above example,

$$\sigma_{est} = \sqrt{\frac{84.65}{8}} = 3.2529$$