Mathematical Representation of Linear Regression, Errors and Residuals

quantra

Contents

In this unit, we will try to understand the mathematical concepts behind the video lectures.

We will understand these concepts by solving simple examples on:

1. Covariance and Correlation
2. Slope and intercept terms from covariance and correlation
3. Interpretation of regression coefficients of the linear regression equation
4. Predicting dependent variable using linear regression equation
5. ANOVA Table analysis
6. How to calculate R^2 and SEE from the ANOVA table

Let us begin by solving for:

Covariance and correlation:

Let 'X' be the returns on S&P 500 and 'Y' be the returns on Stock ABC. Following is the table with observations of X and Y. (A value of 0.12 means 12% return)

Observations (i^{th} value)	X S&P 500 returns	Y Stock ABC returns	$X - X_{mean}$	$(X - X_{mean})^2$	$Y - Y_{mean}$	$(Y - Y_{mean})^2$	$(X - X_{mean})(Y - Y_{mean})$
1	0.12	0.08	0.041	0.00168	0.003	0.00001	0.000123
2	0.34	0.12	0.261	0.06812	0.043	0.00185	0.011223
3	0.06	0.02	-0.019	0.00036	-0.057	0.00325	0.001083
4	0.08	0.1	0.001	0.00000	0.023	0.00053	0.000023
5	0.16	0.25	0.081	0.00656	0.173	0.02993	0.014013
6	-0.13	-0.2	-0.209	0.04368	-0.277	0.07673	0.057893
7	0.02	0.06	-0.059	0.00348	-0.017	0.00029	0.001003
8	0.40	0.19	0.321	0.10304	0.113	0.01277	0.036273
9	-0.21	0.1	-0.289	0.08352	0.023	0.00053	-0.00665
10	-0.05	0.05	-0.129	0.01664	-0.027	0.00073	0.003483

Σ	0.79	0.77	0	0.32708	0	0.12661	0.11847

$\mathbf{X_{mean}}$ = 0.79/10 = 0.079

$\mathbf{Y_{mean}}$ = 0.77/10 = 0.077

Variance of x = s_x^2 = $\Sigma(X - X_{mean})^2/(n-1)$ = 0.32708/9 = 0.03634;

Standard deviation of x = s_x = 0.19063

Variance of y = s_y^2 = $\Sigma(Y - Y_{mean})^2/(n-1)$ = 0.12661/9 = 0.01407;

Standard deviation of y = s_y = 0.11862

Covariance (x,y) = $Cov_{x,y}$ = $\Sigma(X - X_{mean})(Y - Y_{mean})/(n-1)$ = 0.11847/9 **= 0.013163**

Correlation coefficient (x,y) = $r_{x,y}$ = $Cov_{x,y} / s_x * s_y$ = 0.013163/(0.19063)(0.11862) **= 0.58211**

From the above example, we conclude that there is a **positive linear relationship** between S&P 500 and Stock ABC returns since the correlation coefficient is greater than 0 (i.e. $r_{x,y}$ = 0.58211)

Computing slope and intercept terms from covariance and correlation

Further, let us learn the formulae for slope coefficient and intercept given a linear regression equation: $Y_i = b_o + b_1 X_i$

The slope coefficient is: $\mathbf{b_1 = Cov_{x,y} / s_x^2}$

The intercept is: $\mathbf{b_0 = Y_{mean} - b_1 X_{mean}}$

Compute the slope coefficient and intercept of the S&P 500 and Stock ABC example.

We have,

$Cov_{x,y} = 0.013163$

$s_x^2 = 0.03634$

$X_{mean} = 0.079$

$Y_{mean} = 0.077$

Hence the slope coefficient $b_1 = 0.013163/0.03634 = 0.36222$

And the intercept coefficient $b_0 = 0.077 - (0.36222)(0.079) = 0.04838$ i.e. 4.838 %

Interpreting regression coefficients (Slope and Intercept)

Using the above calculated values of S&P 500 and Stock ABC, the linear regression equation looks like:

$Y_i = 0.04838 + 0.36222 * X_i$

In the above equation, the slope coefficient of 0.36222 can be interpreted as when S&P 500 returns increase or decrease by 1%, the corresponding return on Stock ABC would also increase or decrease by 0.36%.

Hence, slope coefficient is called the Stocks ABC's **beta** as explained in the videos.

The intercept term of 4.838% can be interpreted as when the return on S&P 500 is 0%, the Stock ABC's return would be 4.84%.

Hence, intercept is also called Stock ABC's **ex-post alpha** as explained in the video lecture.

Predicting dependent variable using linear regression equation

Once the slope and intercept terms are calculated, we can now predict the Y (Stock ABC's returns) values for a simple linear regression equation.

$\mathbf{Y_{predict} = b_o + b_1 X_i}$

If the returns on S&P 500 are forecasted to be 3% then the return on the Stock ABC would be:

$Y_{ABC} = 0.04838 + 0.36222 * (0.03)$

$Y_{ABC} = 0.05925$

In other words, **the predicted returns on stock ABC is 5.925% when S&P gives a return of 3%**.

ANOVA table (Analysis of Variance)

ANOVA is a statistical procedure for analysing the total variability of the dependent variable. We have already studied a few necessary terms to understand ANOVA table in our video lectures, they are:

i. Total Sum of Squares or SST = $\Sigma (Y_{actual} – Y_{mean})^2$

ii. Regression Sum of Squares or RSS = $\Sigma (Y_{predict} - Y_{mean})^2$

iii. Sum of Squared Errors or SSE = $\Sigma (Y_{actual} - Y_{predict})^2$

iv. Total variation = explained variation + unexplained variation

SST = RSS + SSE

ANOVA table is nothing but summary of the variation in the dependent variable explained above using different terms. A generic ANOVA table for simple linear regression would look like:

Source of Variation	Degree of Freedom	Sum of Squares	Mean Sum of squares
Regression (explained)	1	RSS	MSR = RSS/k = RSS/1 = RSS
Error (Unexplained)	n - k – 1 = n - 2	SSE	MSE = SSE/(n-k-1) = SSE/n-2
Total	n-1	SST	

Where,

'n' is the number of observations

'k' is the number of slope parameters estimated. Because we are limited in simple linear regression (univariate linear regression) our 'k' is always 1.'

Calculate R^2 and SEE from ANOVA table

Question - Complete the ANOVA table for Stock ABC and S&P 500 regression example and calculate the R^2 and the standard error estimate (SEE)

The number of observations is 10.

Source of Variation	Degree of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	?	RSS = 0.00756	?
Error (unexplained)	?	SSE = 0.04064	?
Total	?	?	

Solution: If we fill the ANOVA table with according to what we have learnt:

Source of Variation	Degree of Freedom	Sum of Squares	Mean Sum of Squares
Regression (explained)	1	RSS = 0.00756	MSR = RSS/1 = 0.00756
Error (unexplained)	10 - 2 = 8	SSE = 0.04064	MSE = SSE/8 = 0.00508
Total	9	SST = 0.04820	

R^2 = Explained Variation (RSS) / Total Variation (SST) = **0.15685**

Now since R^2 is 0.15685, it is interpreted as SPX returns explains approximately 15.685% of the variation in return of the Stock ABC over the period.

SEE = $(MSE)^{1/2}$ = $(0.00508)^{1/2}$ = **0.07127**

SEE is the standard deviation of error terms. A value of 0.07127 or 7.127% is not very low. SEE has to be low in order to make strong predictions.

The main reason why R^2 and SEE have not come favourable is because the data set for regression analysis is very small of 10 observations. As we increase the observations, statistically, we will be able to predict the values better.

Ten observations were taken so that you understand the calculations behind simple linear regression. Let us proceed further to study Cost function and Gradient Descent in this section.