R04 Introduction to Linear Regression

Required disclaimer: CFA Institute does not endorse, promote, or warrant the accuracy or quality of the products or services offered by IFT. CFA Institute, CFA®, and Chartered Financial Analyst® are trademarks owned by CFA Institute.

1. Introduction

Financial analysts often need to predict whether one variable X can be used to predict another variable Y. Linear regression allows us to examine this relationship. This reading covers regression analysis with a single independent variable, X_1. In the next reading we will cover regression analysis with multiple independent variables, X_1, X_2, X_3….

2. Linear Regression

2.1. Linear Regression with One Independent Variable

Linear regression assumes a linear relationship between the dependent and independent variables. Regression analysis uses the historical relationship between the independent variable and the dependent variable to predict the values of the dependent variable. The regression equation is expressed as follows:

$$Y_i = b_0 + b_i X_i + \varepsilon_i$$

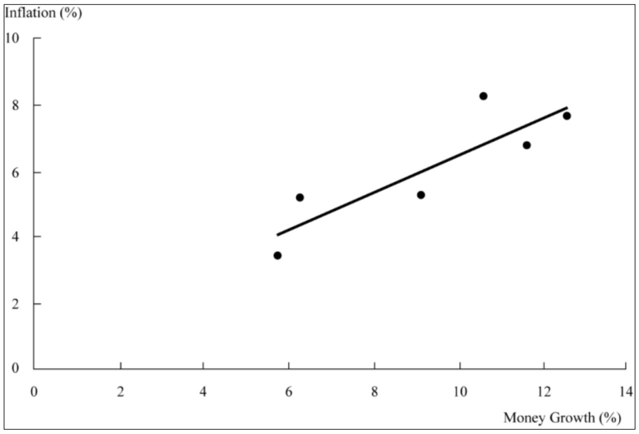
where:
$i = 1, \ldots, n$
Y = dependent variable
b_0 = intercept
b_1 = slope coefficient
X = independent variable
ε = error term
b_0 and b_1 are called the **regression coefficients.**

Dependent variable is the variable being predicted. It is denoted by Y in the equation. The variable used to explain changes in the dependent variable is the independent variable. It is denoted by X. The equation shows how much Y changes when X changes by one unit.

Say that we want to estimate the regression relation between the annual rate of inflation (the dependent variable) and annual rate of money supply growth (the independent variable) for six industrialized countries. The following table shows the data from these countries from 1980 to 2016.

Country	**Money Supply Growth Rate (%)**	**Inflation Rate (%)**
Australia	10.68	4.14
Japan	4.00	0.32
South Korea	16.65	4.86
Switzerland	5.54	1.67
United Kingdom	11.81	4.10
United States	6.28	2.77
Average	9.16	2.98

The figure below demonstrates how linear regression works. It shows a scatter plot constructed using the data for each country. A **linear regression** model or **linear least squares** method, computes the best-fit line through the scatter plot, or the line with the smallest distance between itself and each point on the scatter plot. The regression line may pass through some points, but not through all of them. The vertical distances between the observations and the regression line are called error terms, denoted by $\varepsilon_{i.}$



Linear regression chooses the estimated values for intercept $\widehat{b_0}$ and slope $\widehat{b_1}$ such that the sum of the squared errors (SSE), i.e. the vertical distances between the observations and the regression line is minimized. This is represented by the following equation. The error terms are squared, so that they don't cancel out each other. The objective of the model is that the sum of the squared error terms should be minimized.

$$\sum_{i=1}^{N} \left(Y_i - \widehat{b_0} - \widehat{b_1}X_i\right)^2$$

where $\widehat{b_0}$ and $\widehat{b_1}$ are estimated parameters.

Let's say that the regression software gives us an intercept coefficient of -0.0008 and a slope coefficient of 0.3339. The intercept coefficient tells us that if the money supply growth rate is 0%, then the inflation will be -0.0008%. The slope coefficient tells us that if money supply increases by 1%, then the inflation will increase by 0.3339%. Based on these two coefficients we can come up with the regression equation:

$$\hat{Y} = -0.0008 + 0.3339X$$

This equation can be used to predict inflation rates given money supply growth rate. For example, given a money supply growth rate of 10%, the predicted inflation rate can be calculated as:

$$\hat{Y} = -0.0008 + 0.3339(10\%) = 3.34\%$$

Sample Covariance and Sample Variances

In a regression with one independent variable, the slope coefficient can be expressed as:
Slope coefficient = Cov(Y,X) / Var(X)

If the Cov(Y,X) = 0.0007577 and Var (X) = 0.0022691, the slope coefficient $\hat{b}_1$ = 0.0007577 / 0.0022691 = 0.3339

Note: On the exam you will most likely be given the values of covariance and variance. It is unlikely that you will be asked to calculate these values. Nevertheless, the following table shows how to calculate covariance and variances.

Country	***X_i***	***Y_i***	**Cross-Product**	**Squared Deviations (X)**	**Squared Deviations (Y)**
Australia	0.1068	0.0414	0.0001763	0.0002310	0.0001346
Japan	0.0400	0.0032	0.0013726	0.0026626	0.0007076
South Korea	0.1665	0.0486	0.0014081	0.0056100	0.0003534
Switzerland	0.0554	0.0167	0.0004742	0.0013104	0.0001716
United Kingdom	0.1181	0.0410	0.0002968	0.0007023	0.0001254
United States	0.0628	0.0277	0.0000605	0.0008294	0.0000044
Sum	0.5496	0.1786	0.0037885	0.0113457	0.0014970
Average	0.0916	0.0298			
Covariance			0.0007577		
Variance				0.0022691	0.0002994
Standard deviation				0.0476356	0.0173033

Notes:

1. Divide the cross-product sum by $n - 1$ (with $n = 6$) to obtain the covariance of X and Y.
2. Divide the squared deviations sums by $n - 1$ (with $n = 6$) to obtain the variances of X and Y.

3. Assumptions of the Linear Regression Model

The classic linear regression model ($Y_i = b_0 + b_1 Xi + \varepsilon_i$ where i = 1, …, n) is based on the following six assumptions:

1. The relationship between Y and X is linear in the parameters b_0 and b_1. This means that b_0, b_1, and X can only be raised to the power of 1. b_0 and b_1 cannot be multiplied or divided by another regression parameter. This assumption is critical for a valid linear regression.
2. X is not random.
3. The expected value of the error term is zero. This ensures the model produces correct estimates for b_0 and b_1.
4. The variance of the error term is constant for all observations. This is called homoskedasticity. If the variance of the error term is not constant, then it is called heteroskedasticity.
5. The error term, ε, is uncorrelated across observations.
6. The error term, ε, is normally distributed.

4. The Standard Error of Estimate

The standard error of estimate (SEE) measures how well a given linear regression model captures the relationship between the dependent and independent variables. It is the standard deviation of the prediction errors. A low SEE implies an accurate forecast.

The formula for the standard error of estimate is given below:

$$\text{SEE} = \left(\sum_{i=1}^{N} \frac{\left(Y_i - \widehat{b_0} - \widehat{b_1} X_i\right)^2}{n - 2} \right)^{\frac{1}{2}} = \left(\frac{\sum_{i=1}^{N} \left(\widehat{\varepsilon_i}\right)^2}{n - 2} \right)^{\frac{1}{2}}$$

A low SEE implies that the error (or residual) terms are small and hence the linear regression model does a good job of capturing the relationship between dependent and independent variables.

5. The Coefficient of Determination

The coefficient of determination, denoted by R^2, measures the fraction of the total variation in the dependent variable that is explained by the independent variable.

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$$

Total variation = $\sum_{i=1}^{N} (y_i - \bar{y})^2$

Unexplained variation = $\sum_{i=1}^{N} (y_i - \hat{y})^2$

where:
y_i = actual value of the dependent variable for a given value of the independent variable.

$\overline{Y}$= the average value of y_i.

Characteristics of coefficient of determination, R^2

The higher the R^2, the more useful the model. R^2 has a value between 0 and 1. For example, if R^2 is 0.03, then this explains only 3% of the variation in the independent variable and the explanatory power of the model is low. However, if R^2 is 0.85, this explains over 85% of the variation and the explanatory power of the model is high.

It tells us how much better the prediction is by using the regression equation rather than just $\bar{y}$ (average value) to predict y.

With only one independent variable, R^2 is the square of the correlation.

The correlation, r, is also called the "multiple - R".

6. Hypothesis Testing

In order to test whether an estimated regression coefficient is statistically significant, we use hypothesis testing. There are two ways to perform hypothesis testing using either (1) the confidence interval or (2) testing for a significance level, that we will see in the following example.

Suppose we regress a stock's returns on a stock market index's returns and find that the slope coefficient is 1.5 with a standard error estimate of 0.200. Assume we need 60 monthly observations in our regression analysis. The hypothesized value of the parameter is 1.0, which is the market average slope coefficient. Define the null and alternative hypothesis. Compute the test statistic. At a 0.05 level of significance, should we reject H_0?

Solution:

Method 1: The confidence interval approach

Step 1: Select the significance level for the test. For instance, if we are testing at a significance level of 0.05, then we will construct a 95 percent confidence interval.

Step 2: Determine the critical t value t_c for a given level of significance and degrees of freedom. In this regression with one independent variable, there are two estimated parameters, the intercept and the coefficient on the independent variable.

Number of degrees of freedom = number of observations – number of estimated parameters = 60 – 2 = 58.
t_c = 2.00 at (significance level of 0.05 ad 58 degrees of freedom)

Step 3: Construct the confidence interval which will span from $\widehat{b_1} - t_c s_{\widehat{b_1}}$ to $\widehat{b_1} + t_c s_{\widehat{b_1}}$ where t_c is the critical t value.
Confidence interval = 1.5 – 2(0.2) to 1.5 + 2(0.2) = 1.10 to 1.90

Step 4: Conclusion: Make a decision to reject or fail to reject the null hypothesis.
Since the hypothesized value of 1.0 for the slope coefficient falls outside the confidence interval, we can reject the null hypothesis. This means we are 95 percent confident that the interval for the slope coefficient does not contain 1.0.

Method 2: Using the t-test of significance

Step 1: Select the significance level for the test. Here we select a significance level of 5%.

Step 2: Define the null and alternate hypothesis. H_0: $b_1 = 1.0$; $H_{a:}$ $b_1 \neq 1.0$

Step 3: Compute the t-statistic using the formula below:

$$t = \frac{\widehat{b_1} - b_1}{s_{\widehat{b_1}}}$$

The t-statistic is for a t-distribution with (n - 2) degrees of freedom because two parameters are estimated in the regression.

$$t = \frac{1.5 - 1.0}{0.2} = 2.5$$

Step 4: Compare the absolute value of the t-statistic to $t_{c.}$ and make a decision to reject or fail to reject the null hypothesis. If the absolute value of t is greater than t_c, then reject the null hypothesis. Since 2.5 > 2, we can reject the null hypothesis that $b_1 = 1.0$.

Notice that both the approaches give the same result.

p-value: At times financial analysts report the *p*-value or probability value for a particular hypothesis. The p-value is the smallest level of significance at which the null hypothesis can be rejected. It allows the reader to interpret the results rather than be told that a certain hypothesis has been rejected or accepted. In most regression software packages, the p-values printed for regression coefficients apply to a test of the null hypothesis that the true parameter is equal to 0 against the alternative that the parameter is not equal to 0, given the estimated coefficient and the standard error for that coefficient. Here are a few important points connecting t-statistic and p-value:

- Higher the t-statistic, smaller the p-value.
- The smaller the p-value, the stronger the evidence to reject the null hypothesis.
- Given a p-value, if p-value ≤ α, then reject the null hypothesis H_0. α is the significance level.

7. Analysis of Variance in a Regression with One Independent Variable

Analysis of variance or ANOVA is a statistical procedure of dividing the total variability of a variable into components that can be attributed to different sources. We use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable. In simple terms, ANOVA explains the variation in the dependent variable for different levels of the independent variables.

For a meaningful regression model the slope coefficients should be non-zero. This is determined through the **F-test** which is based on the **F-statistic**. The F-statistic tests whether **<u>all</u>** the slope coefficients in a linear regression are equal to 0. In a regression with one independent variable, this is a test of the null hypothesis $H_{0:}$ $b_1 = 0$ against the alternative hypothesis H_a: $b_1 \neq 0$. The F-statistic also measures how well the regression equation explains the variation in the dependent variable. The four values required to construct the F-statistic for null hypothesis testing are:

- The total number of observations (n)
- The total number of parameters to be estimated
- The sum of squared errors or residuals (SSE): $\sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2$. This is also known as residual sum of squares.
- The regression sum of squares (RSS), $\sum_{i=1}^{N} (\hat{Y}_i - \bar{Y})^2$. This is the amount of total variation in Y that is explained in the regression equation.

The F-statistic is the ratio of the average regression sum of squares to the average sum of the squared errors. Average regression sum of squares (RSS) is the amount of variation in Y explained by the regression equation.

$$F = \frac{\frac{RSS}{1}}{\frac{SSE}{n-2}} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}}$$

where:

$$\text{Mean regression sum of squares} = \frac{\text{regression sum of squares}}{\text{number of slope parameters estimated}}$$

$$\text{Mean squared error} = \frac{\text{sum of squared errors}}{\text{number of observations n} - \text{ total number of parameters estimated}}$$

Interpretation of F-test statistic:

- The higher the F-statistic, the better.
- A high F-statistic implies that the regression model does a good job of explaining the variation in the dependent variable.
- A low F-statistic implies that the regression model does not do a good job of explaining the variation in the dependent variable.
- A F-statistic of 0 indicates that the independent variable does not explain variation in the dependent variable.

F-statistic is usually not used in a regression with one independent variable because the F-statistic is the square of the t-statistic for the slope coefficient and it tells us the same thing as the t-test of the slope coefficient. This is illustrated in Example 18.

Note: The F-statistic will be covered in more detail in the next reading on multiple regression.

Example: Performance evaluation: The Dreyfus Appreciation Fund

This is based on Example 7 in the curriculum.

This example illustrates how the F-test reveals nothing more than what the t-test already does, when we have just one independent variable. The ANOVA table and results of regression to evaluate the performance of the Dreyfus appreciation fund is given below:

Regression Statistics			
Multiple *R*	0.928		
R-squared	0.8611		
Standard error of estimate	0.0174		
Observations	60		
ANOVA	**Degrees of Freedom (df)**	**Mean Sum of Squares (MSS)**	***F***
Regression	1	0.1093	359.64
Residual	58	0.0003	
Total	59		

	Coefficients	**Standard Error**	***t*-Statistic**
Alpha	0.0009	0.0023	0.4036
Beta	0.7902	0.0417	18.9655

Source: Center for Research in Security Prices, University of Chicago.

Note: From an exam perspective, be familiar with how to read the values from a table like this and understand what values matter in coming to a conclusion.

Test the null hypothesis that α = 0, i.e. the fund did not have a significant excess return beyond the return associated with the market risk of the fund.

Solution:

The software tests the null hypothesis for slope coefficient = 0, so we can use the given t-statistic. Since the t-statistic of 18.9655 is high, the probability that the coefficient is 0 is very small.

We get the same result using the F-statistic as well.

$$F = \frac{\frac{0.1093}{1}}{\frac{0.0176}{60-2}} = 360.19$$

The p-value for F-statistic is less than 0.0001. This is much lower than the significance level.

Additional parameters that can be calculated directly from the ANOVA table:

1. Total variation = explained variation + unexplained variation i.e. SST = RSS + SSE
2. Sample variance = SST/n-1(degrees of freedom)
3. R^2 can be calculated as:
$$R^2 = \frac{SST - SSE}{SST} = \frac{RSS}{SST}$$
4. SEE can be calculated as:
$$SEE = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}}$$

8. Prediction Intervals

We use regression equations to make predictions about a dependent variable. Confidence intervals for the predicted value of a dependent variable are calculated just like the confidence intervals for the regression coefficient.

Let us consider the regression equation: $Y = b_0 + b_1X$. The predicted value of $\widehat{Y} = \widehat{b_0} + \widehat{b_1}X$. The two sources of uncertainty to make a prediction are:

1. The error term
2. Uncertainty in predicting the estimated parameters b_0 and b_1.

The estimated variance of the prediction error is given by:

$$s_f^2 = s^2 * \left[1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1)s_x^2}\right]$$

Note: You need not memorize this formula, but understand the factors that affect s_f^2, like higher the n, lower the variance, and the better it is.

The estimated variance depends on:

- the squared standard error of estimate, s^2
- the number of observations, n
- the value of the independent variable, X
- the estimated mean $\overline{X}$
- variance, s^2, of the independent variable

Once the variance of the prediction error is known, it is easy to determine the confidence interval around the prediction. The steps are:

1. Make the prediction.
2. Compute the variance of the prediction error.
3. Determine t_c at the chosen significance level α.
4. Compute the (1-α) prediction interval using the formula below:
   $\hat{Y} \pm t_c * s_f$

Example: Predicting the ratio of enterprise value to invested capital

This is based on Example 8 in the curriculum.

Given the results of a regression analysis below, predict the ratio of enterprise value to invested capital for the company at a 95 percent confidence interval. The return spread between ROIC and WACC is given as 10 percentage points.

Regression Statistics			
Multiple *R*	0.9469		
R-squared	0.8966		
Standard error of estimate	0.7422		
Observations	9		
	Coefficients	**Standard Error**	***t*-Statistic**
Intercept	1.3478	0.3511	3.8391
Spread	30.0169	3.8519	7.7928

Source: Nelson (2003).

Solution:

1. Expected EV/IC = $b_0 + b_1$ (ROIC-WACC) + ε

Plugging in values from the table, we have:

Expected EV/IC = 1.3478 + 30.0169 (0.1) = 4.3495

This means that if the return spread is 10 percent, then the EV/IC ratio will be 4.3495.

2. Compute the variance of the prediction error.

$$s_f^2 = s^2 * \left[1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{(n-1)s_x^2}\right] = 0.7422^2 * \left[1 + \frac{1}{9} + \frac{(0.1 - 0.0647)^2}{(9-1)0.004641}\right] = 0.630556$$

S_f = 0.7941

3. t_c at the 95 percent confidence level and 7 degrees of freedom is 2.365.

4. 95 percent prediction interval using equation 8 is:
4.3495– 2.365 (0.7941) to 4.3495 + 2.365 (0.7941) = 2.4715 to 6.2275.

9. Limitations of Regression Analysis

Regression relations can change over time as do correlations. This is called parameter instability. This characteristic is observed in both cross-series and time-series regression relationships.

Regression analysis is often difficult to apply because of specification issues and assumption violations.

- Specification issues: Identifying the independent variables and the dependent variable, and formulating the regression equation may be challenging.
- Assumptions violations: Often there is an uncertainty in determining if an underlying assumption has been violated.

Public knowledge of regression relationships may limit their usefulness in the future. For example, is low P/E stocks in the sugar industry have had historically high returns during Oct-Jan every year, then this knowledge may cause other analysts to buy the stock which will push their prices up.

Summary

LO.a: Distinguish between the dependent and independent variables in a linear regression.

Linear regression assumes a linear relationship between the dependent variable and the independent variable.

The variable used to explain the dependent variable is the independent variable, X. The variable being explained is the dependent variable, Y.

A simple linear regression using one independent variable can be expressed as:

$Y_i = b_0 + b_iX_i + \varepsilon_i$ where i = 1,2,…n

LO.b: Explain the assumptions underlying linear regression and interpret regression coefficients.

1. The relationship between Y and X is linear.
2. X is not random.
3. The expected value of the error term is zero.
4. The variance of the error term is constant for all observations.
5. The error term, ε, is uncorrelated across observations.
6. The error term, ε, is normally distributed.

LO.c: Calculate and interpret the standard error of estimate, the coefficient of determination, and a confidence interval for a regression coefficient.

- Standard error of estimate: $\mathrm{SEE} = \left(\sum_{i=1}^{N} \frac{(Y_i - \widehat{b_0} - \widehat{b_1}X_i)^2}{n-2}\right)^{\frac{1}{2}} = \left(\frac{\sum_{i=1}^{N}(\widehat{\varepsilon_i})^2}{n-2}\right)^{\frac{1}{2}}$
- The lower the SEE, the better the fit of the regression line.
- Coefficient of determination $R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{\text{Total variation} - \text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}$
- The higher the R^2, the more useful the model. R^2 has a value between 0 and 1.
- The confidence interval for a regression coefficient is given by $\widehat{b_1} - t_c s_{\widehat{b_1}}$ to $\widehat{b_1} + t_c s_{\widehat{b_1}}$ where t_c is the critical t value for a given level of significance and degrees of freedom and $\widehat{b_1}$ is the estimated slope coefficient. If the hypothesized value is outside the confidence interval, then reject the null hypothesis.

LO.d: Formulate a null and alternative hypothesis about a population value of a regression coefficient and determine the appropriate test statistic and whether the null hypothesis is rejected at a given level of significance.

- Compare t-statistic = t = $\frac{\widehat{b_1} - b_1}{s_{\widehat{b_1}}}$with t_c; if t > $t_{c,}$ then reject the null.
- The p-value is the smallest level of significance at which the null hypothesis can be rejected. Higher the t-statistic, smaller the p-value and the stronger the evidence to reject the null hypothesis.

LO.e: Calculate the predicted value for the dependent variable, given an estimated regression model and a value for the independent variable.

The predicted value of the dependent variable(Y) is calculated by inserting the predicted value of the independent variable(X) in the regression equation.

$\mathrm{Y_i = b_0 + b_i X_i}$

LO.f: Calculate and interpret a confidence interval for the predicted value of the dependent variable.

The prediction interval is given by: $\hat{\mathrm{Y}} \pm \mathrm{t_c} * \mathrm{s_f}$

LO.g: Describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the F-statistic.

- Analysis of variance is a statistical procedure for dividing the variability of a variable into components that can be attributed to different sources. We use ANOVA to determine the usefulness of the independent variable or variables in explaining variation in the dependent variable.
- The F-statistic measures how well the regression equation explains the variation in the dependent variable.

$$\mathrm{F} = \frac{\frac{\mathrm{RSS}}{1}}{\frac{\mathrm{SSE}}{\mathrm{n-2}}} = \frac{\text{Mean regression sum of squares}}{\text{Mean squared error}}$$

where:

Mean regression sum of squares = $\frac{\text{regression sum of squares}}{\text{number of slope parameters estimated}}$

Mean squared error = $\frac{\text{sum of squared errors}}{\text{number of observations n – total number of parameters estimated}}$

LO.h: Describe limitations of regression analysis.

- Regression relations can change over time as do correlations. This is called parameter instability.
- Regression analysis is often difficult to apply because of specification issues and assumption violations.
- Public knowledge of regression relationships may limit their usefulness in the future.