Imperial College London

MSc Financial Technology

BUSI70606 Financial Econometrics

Tutorials

Module Leader:
Roald Versteeg

Teaching Assistant:
Luca Luigi Alberici

October 3, 2025

Problem Set 1 - Regression Analysis

1. Explain with the use of equations, the difference between the sample regression function and the population regression function.

2. What five assumptions are usually made about the unobservable error terms in the classical linear regression model (CLRM)? Briefly explain the meaning of each. Why are these assumptions made?

3. Which of the following models can be estimated (following a suitable rearrangement if necessary) using ordinary least squares (OLS)?
   x, y, z are variables and α, β, γ are parameters to be estimated.

   (a) $y_t = \alpha + \beta x_t + u_t$
   (b) $y_t = e^{\beta t} x_t^{\beta} e^{u_t}$
   (c) $y_t = \alpha + \beta\gamma x_t + u_t$
   (d) $\ln(y_t) = \alpha + \beta \ln(x_t) + u_t$
   (e) $y_t = \alpha + \beta x_t z_t + u_t$

4. To estimate the CAPM beta of a stock one can run the regression:

$$[R_{it} - R_{rft}] = \alpha_i + \beta_i [R_{mt} - R_{rft}] + u_{it},$$

   where $[E(R_{it}) - R_{ft}]$ is the excess return on stock i and where $[E(R_{mt}) - R_{rft}]$ is the excess return on the market.

   (a) Assume that, using 62 observations, you have estimated a beta of 1.147 (with a standard error of 0.0548) for IBM. Test, at the 5% level, the null hypothesis that IBM is as risky (no more no less) than the market. Test this null against the single sided hypothesis that IBM is more risky than the market.

   (b) Now assume that, using 38 observations, you have estimated a beta of 0.214 (with a standard error of 0.186) for Acorn Mining. Test, at the 5% level, the null hypothesis that Acorn's returns do not have any systematic risk.
   (In other words the correlation between Acorn's returns and market returns is zero). Test this null against a two-sided alternative.

   (c) Form and interpret a 95% and 99% confidence interval for the beta you calculated in 4b.

5. Are hypothesis tested concerning the actual values of the coefficients (i.e. β) or their estimated values (i.e. $\hat{\beta}$)?

PS1 [Solution] - Regression Analysis

1. The population regression function is a linear function that describes the true relationship in the population between the two variables - how the observed values of X are related to the observed values of Y exactly. You can also consider this a different way: if the variable Y is 'generated' in the real world through a data-generating process that involves X, the PRF describes this using the *true* values of α and β.

$$Y_t = \alpha + \beta X_t + u_t$$

Note that the Y_t here is the real observed value of Y in the population, not a construction or estimate. The existence of u_t here is therefore necessary, because we usually don't think of Y_t and X_t as perfectly linearly related. u_t captures measurement errors, omitted variables, and randomness as a result of factors not accounted for by X.

The sample regression function is the estimate of the relationship between X and Y based on the **sample**. We form an estimate of the parameters of the population regression function (α and β) using our sample.

$$\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$$

This uses the values of X that are actually observed (in the sample), but $\hat{Y}_t$ depends on $\hat{\alpha}$ and $\hat{\beta}$. Note that there is no error here. That's because $\hat{Y}_t$ is an approximation of Y_t based on the line of best fit: this equation just states that a linear function of observed X gives a certain estimated value of Y, that is $\hat{Y}$.

$$Y_t = \hat{\alpha} + \hat{\beta} X_t + \hat{u}_t$$

This is not the SRF, but it explains that the true observed value of Y_t can be split into the fitted value from the model $\hat{Y}$, and a residual $\hat{u}_t$. The SRF is thus used to infer the likeliest values of the parameters of the PRF, using $\hat{\alpha}$ and $\hat{\beta}$.

2. The assumptions of the CLRM model, otherwise the Gauss-Markov assumptions are as follows. These can also be found in Box 3.3. of Brooks' book.

   - $E(u_t) = 0$; the errors have 0 mean. If this assumption is violated, this means that there is some variation that the model doesn't yet capture.
   - $var(u_t) = \sigma^2 < \infty \; \forall \; x_t$; the variance of the errors is constant and independent of x_t.
   - $cov(u_i, u_j) = 0$; errors are linearly independent of each other.
   - $cov(u_t, x_t) = 0$; errors are linearly independent of the explanatory variable.
   - u_t is normally distributed.

   Together, the first four assumptions prove that the OLS estimators (α and β) are the "best" among linear unbiased estimators (BLUE), in that they have the minimum variance of the class of linear unbiased estimators. The theorem that the OLS estimators are BLUE is called the **Gauss-Markov theorem**. Violations of these assumptions create situations in which the OLS is no longer the best option, due to either bias or efficiency loss (or both). Therefore, they may be inaccurate with regard to the relationship between variables, and subject to fluctuations between samples.

The fifth assumption is useful because it enables us to make statistical inferences about the population parameters from the sample data - in other words, to test hypotheses about the coefficients. This assumption, provided the others hold, implies that the test statistics follow a **t-distribution**, the basis of hypothesis testing.

3.

$$y_t = \alpha + \beta x_t + u_t$$

Yes, we can use OLS. This is the usual univariate linear model we have been dealing with.

$$y_t = e^{\alpha} x_t^{\beta} e^{u_t}$$

Yes, the model can be linearised by taking logarithms of both sides and by rearranging.

$$ln(y_t) = \alpha + \beta ln(x_t) + u_t$$

After transforming one's data using the natural log, we can estimate the above. Several theories make use of this form: for example, Cobb-Douglas production functions in economics. Logarithmic transformations are thus a useful way to find a linear approximation of non-linear relationships to estimate them. Log transformations can also reduce the effect of extreme values in the data and thereby reduce skewness. Using logarithms can also turn multiplicative models into additive ones which we can easily estimate.

The interpretation of coefficients in words will change depending on whether you use log or not, and where you use log (in the independent variable, in the dependent variable, or both).

$$y_t = \alpha + \beta\gamma x_t + u_t$$

We can use OLS to estimate this model, but we would not be able to obtain the values of β and γ independently, only their product.

$$ln(y_t) = \alpha + \beta ln(x_t) + u_t$$

Yes, we can use OLS. This model is linear in the logarithms, and after transforming our data into logs, we can run this regression. Those with some background in economics may recognise that α and β in this case can be interpreted as elasticities.

$$y_t = \alpha + \beta x_t z_t + u_t$$

Yes, we can use OLS as it is linear in the parameters. We would have to construct a new variable that is the product of x_t and z_t for all t, which we can name q_t. Then we can run the regression $y_t = \alpha + \beta q_t + u_t$ as usual. We can estimate a fairly wide range of model types using these simple tools.

4. • The null hypothesis that IBM is as risky as the market implies that beta = 1.

$$H_0 : \beta = 1$$

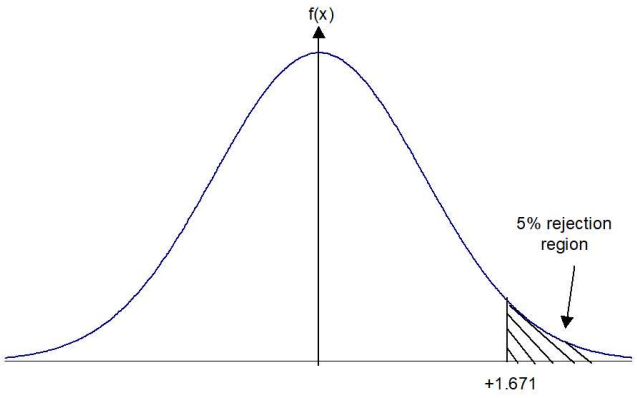
Testing against the hypothesis that the fund is riskier than the market means that our alternate hypothesis is

$$H_1 : \beta > 1$$

The next step, calculate your t-statistic.

$$test\ stat = \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})} = \frac{1.147 - 1}{0.0548} = 2.682$$

We want to compare our test statistic with a value from the t-table for $T - 2$ degrees of freedom, where T is the sample size - so 60 degrees of freedom in this case. As we are doing a one-sided test, we want a value with 5% all in one tail – the critical value for this is 1.671.



Since the test statistic is larger than the critical value, we can reject the null hypothesis. We have statistically significant evidence that this security has a beta greater than 1 or in other words, it is more risky than the market as a whole.

- We follow the same steps as before, but this time, we want to test the hypothesis that there's no systematic risk in the shares of Acorn Mining against a two-sided alternative. In other words, the null hypothesis is that value of beta in the regression model is zero, and no matter what happens to the market proxy, Acorn Mining would be completely unaffected by it.

$$H_0 : \beta = 0$$

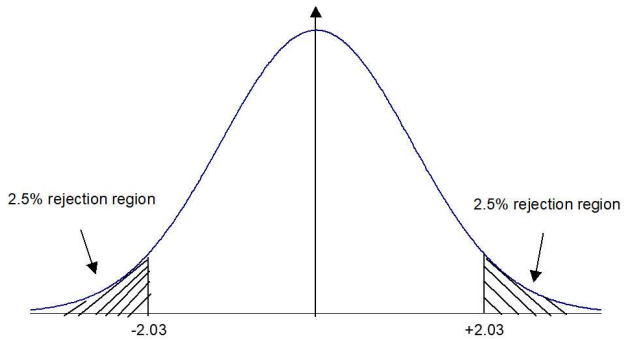
Two-sided means we don't make any prior assessments as to the value of beta - could move in the same direction as the market or the opposite.

$$H_1 : \beta \neq 0$$

Calculate the test statistic.

$$test\ stat = \frac{\hat{\beta} - \beta^*}{SE(\hat{\beta})} = \frac{0.214 - 0}{0.186} = 1.150$$

The sample size is 38, and therefore, the test statistic follows a T-distribution with 36 degrees of freedom. Selecting a 5% significance level in a two-sided test means that we need to find two critical values that together cover 5% of the probability. Therefore, we find two critical values that put 2.5% of the distribution in each tail.



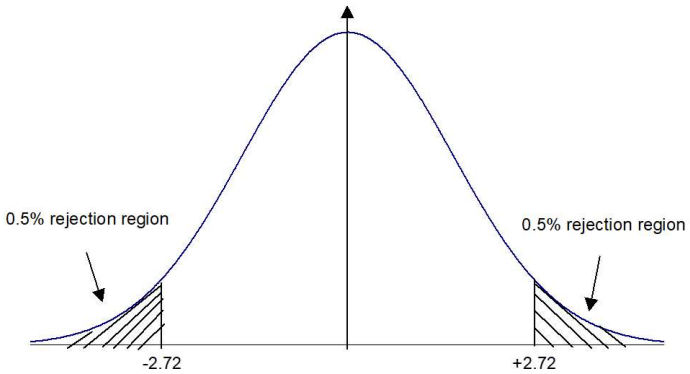
The critical values are -2.03 and +2.03.

The test statistic is not in the rejection region, and therefore, we fail to reject the null hypothesis. There is therefore no statistically significant evidence that Chris Mining has any systematic risk. In other words, we have no evidence that changes in the company's value are driven by movements in the market.

- A confidence interval for the β is an interval defined such that there is a specific probability that the parameter value occurs within the interval. To calculate the confidence interval, we use the formula

$$(\hat{\beta} - SE(\hat{\beta}) \times t_{crit},\ \hat{\beta} + SE(\hat{\beta}) \times t_{crit})$$

Confidence intervals are almost always two sided, unless specified otherwise. Therefore, we look up the values that place 2.5% probability density on the lower tail and the upper tail for a 95% confidence interval, and 0.5% on either tail for 99%.



The confidence intervals are

$$(0.214 \pm 0.186 \times 2.03) = (-0.164, 0.592)$$

$$(0.214 \pm 0.186 \times 2.72) = (-0.292, 0.720)$$

We note two further points about confidence intervals. First, we can intuitively interpret the X% confidence interval as us being X% sure that the true value of the population parameter lies within the interval. So we can be 95% sure that the true value of beta lies within the interval (-0.164, 0.592) and we are 99% sure that the true populationv alue of beta lies within (-0.292, 0.720). Thus, in order to be more sure that we have the true value of beta contained in the interval, the interval must become wider.

Secondly, once we have formed the interval, we can test an infinite number of hypotheses. We failed to reject the null hypothesis in the previous part because 0 exists in the interval at both 95% and 99% confidence. We can reject the null hypothesis that the true value is 0.6 at a 95% confidence level because it's not in the confidence interval corresponding to 95%, but the same cannot be said for 99% confidence. Therefore we should always, if possible, conduct some sort of sensitivity analysis to see if our conclusions are altered by (sensible) changes in the level of significance used.

5. We test hypotheses for on the actual coefficients, not the estimated values. The values of the actual coefficients are unknown, and therefore require inference (hypothesis testing) to find out. We don't need to test for the 'true' values of our estimated coefficients, since we know exactly what they are; we just calculated them!