Module 1, section 2 : Regression, the CAPM and Factor Models

Chapters 2 and 3 in Brooks
Chapter 2 in Tsay
Chapter 5 in Campbell, Lo and Mackinlay

1 Regression

- Regression concerns the relationship between two or more variables. It is a fundamental building block of quantitative finance.

- The purpose is to explain how (the average of) one variable changes, as another variable (or variables) changes.

- Consider the simple linear regression (SLR) model:

$$y_t = \alpha + \beta X_t + \epsilon_t$$

- The conditional expectation of y (dependent) given X (explanatory) is: $E(y_t|X_t) = \mu_t = \alpha + \beta X_t$.

- Most regression and time series models are similarly based on conditional expectations and/or conditional distributions.

- The error term can contain: omitted variables, measurement error, ..., something else??

- What is an omitted variable? Can such omissions negate or make our analysis worthless?

- If we also assumed a constant error variance, i.e. $Var(\epsilon_t) = \sigma^2$, then:

$$E(y_t|X_t) = \alpha + \beta X_t \ ; \ Var(y_t|X_t) = \sigma^2$$

- Clearly, in this case y and X are dependent (not independent), i.e. they have a relationship, since if two variables are independent, then:

$$E(Y|X) = E(Y); Var(Y|X) = Var(Y).$$

- β is the average change in y when X increases by 1 unit. *check this!*.

- The direction of the relationship is assumed to be from the explanatory variable, X, to the dependent or response variable y.

- Regression is often used to assess causality: i.e. do changes in X **cause** changes in y.

- Conditions for **scientific causality**

  1. When X changes Y changes
  2. X occurs before Y occurs

3. No other variable could have caused the observed change in Y

- Are these conditions usually satisfied with real or empirical data?

- Could they ever be satisfied?

- Correlation is another measure of the linear relationship between two variables.

- It does not assume a direction for the relationship, it simply shows ... *what?*.

- Some relevant theory and notation:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

where

$$\mathrm{Cov}(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

and

$$\mathrm{Var}(Y) = E\left[(Y - \mu_Y)^2\right]$$

- Correlation as a measure is:

1. unit and scale free.
2. always between -1 and 1 in value
3. positive if Y and X increase together, on average, and negative otherwise
4. equal to 0 whenever Y and X are independent AND when . . . *?*.
5. equal to 1 or -1 if Y or X can be predicted **exactly** given knowledge of the other, in a **linear** relationship.
6. a measure of the strength ('away' from 0) or weakness ('close' to 0) of the **linear** relationship between Y and X.
7. **irrelevant** if Y and X are NOT linearly related

- Figure 1 highlights some more properties of correlation as a measure.

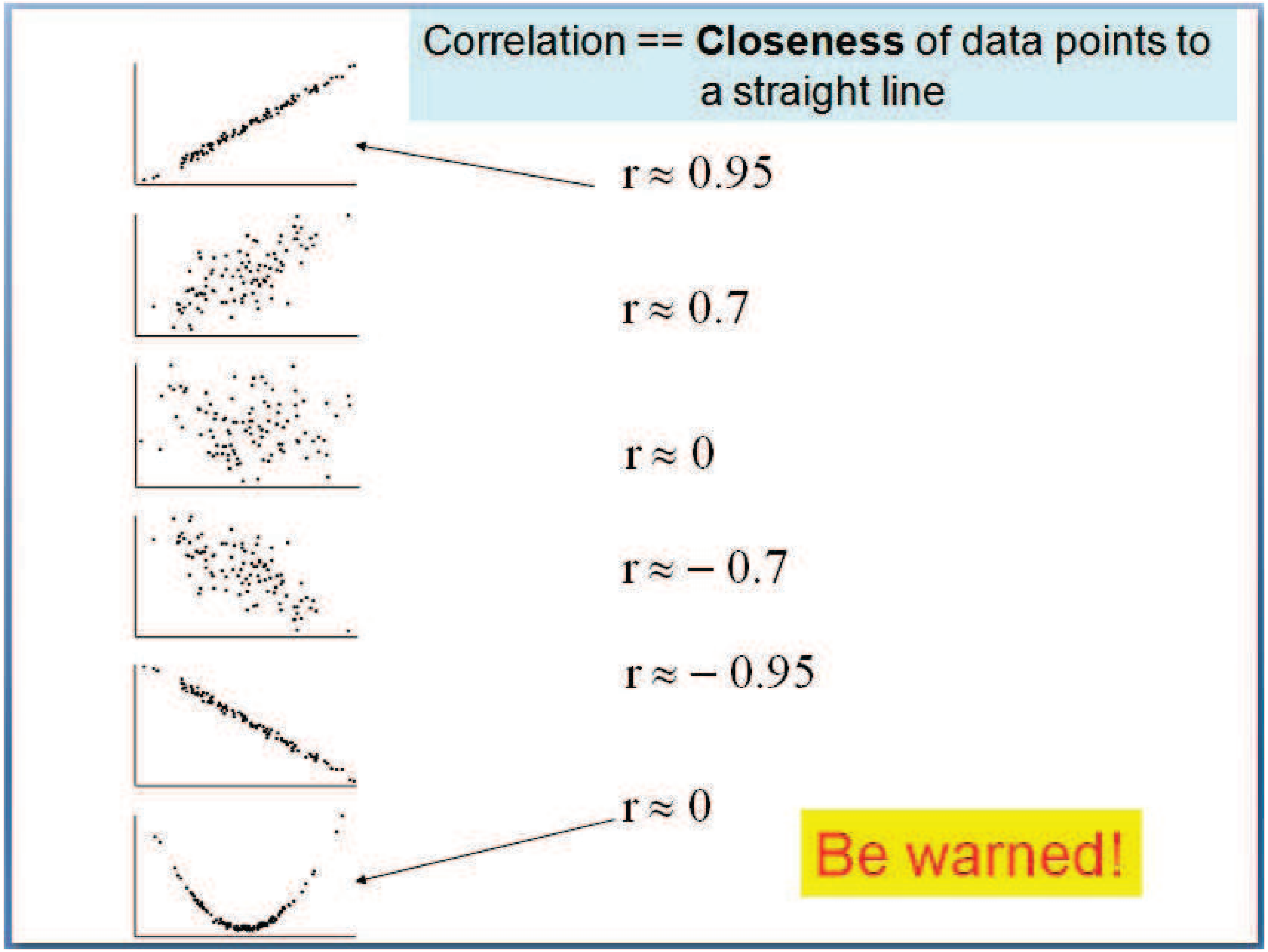

Figure 1: Some data sets with different correlations

- Correlation is usually denoted ρ and either r or $\hat{\rho}$ when estimated. We'll use $\hat{\rho}$ mostly.

- Correlation assumes Y and X have constant, finite unconditional expectations and variances.

- Correlation assumes NO direction or causation in the relationship.

- To estimate correlation from a sample of T observations $(x_1, y_1), \ldots, (x_T, y_T)$, a common estimator is developed as:

$$\widehat{\mathrm{Cov}(X,Y)} = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})$$
$$\widehat{\mathrm{Var}(X)} = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})^2$$

leading to

$$\widehat{\rho} = \frac{\widehat{\mathrm{Cov}}(X,Y)}{\sqrt{\widehat{\mathrm{Var}}(X)\widehat{\mathrm{Var}}(Y)}} = \frac{\sum_{t=1}^{T}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2 \sum_{t=1}^{T}(y_t - \bar{y})^2}}$$

- Why is it common? …Is it sensible?

- A t-test can be done regarding whether the true correlation ρ could be 0 or not:

$$t = \frac{\hat{\rho}\sqrt{T-2}}{\sqrt{1-\hat{\rho}^2}}$$

which has a central limit theorem, where t follows a t-distribution with $T-2$ degrees of freedom.

- The hypotheses are Null: $\rho = 0$ vs Alternative: $\rho \neq 0$.

- The p-value is the probability of observing a value of $\hat{\rho}$ (i.e. t) as far, or further away, from 0 in a sample of size T if $\rho = 0$ was true.

- The test assumes that either T is *large* OR Y and X are normally distributed, or both.

- And it assumes that the sample of data is identically and independently distributed (iid).

- Finally, both Y and X must have finite (1st, 2nd, 3rd and) 4th unconditional moments, for this test to work properly.

- What does all this mean?? Why are these assumptions necessary?

- Figure 2 shows a seemingly innocuous pdf that has an infinite variance.

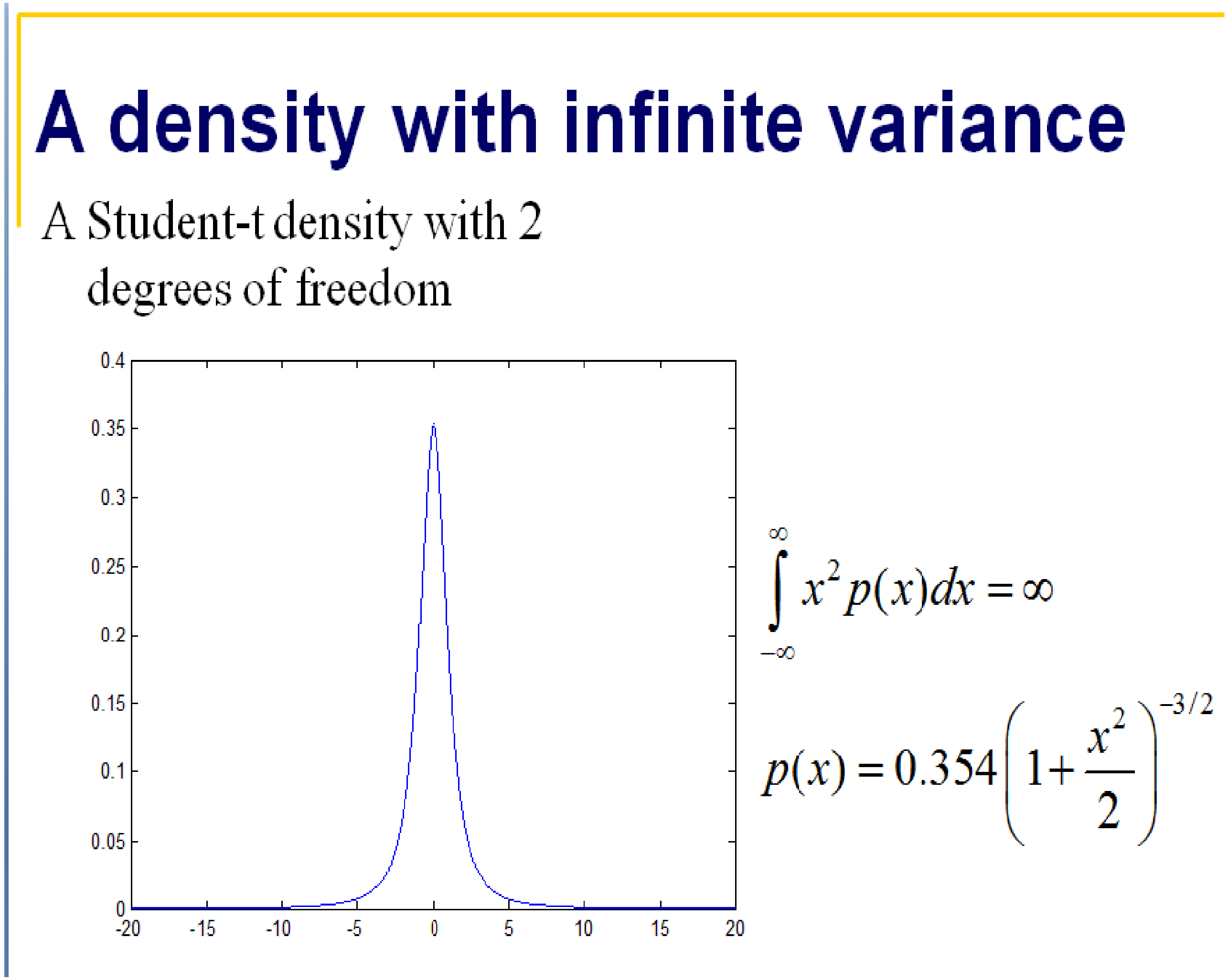

Figure 2: A continuous distribution for an rv with infinite variance

Example

- Figure 3 shows the daily index and log return values for the AORD and S&P500 indices, from January, 2000 until February, 2017.

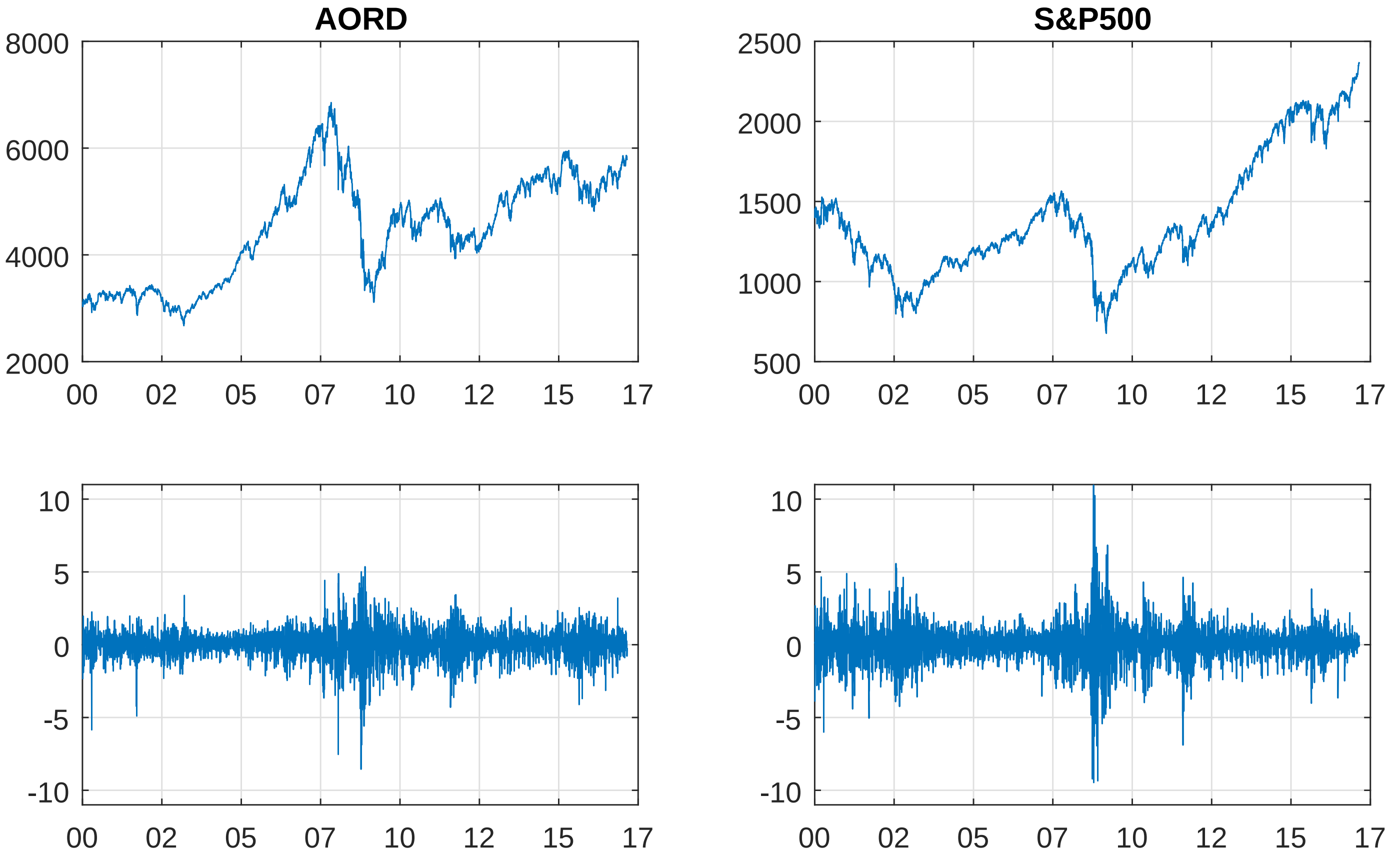

Figure 3: Index values and daily log returns for AORD and SP500 indices

- Question: is there a relationship between AORD and S&P500 daily returns?

- Figure 4 shows a scatterplot of AORD vs S&P500 daily log-returns, each on the same calendar days (i.e. contemporaneous) when both markets traded.

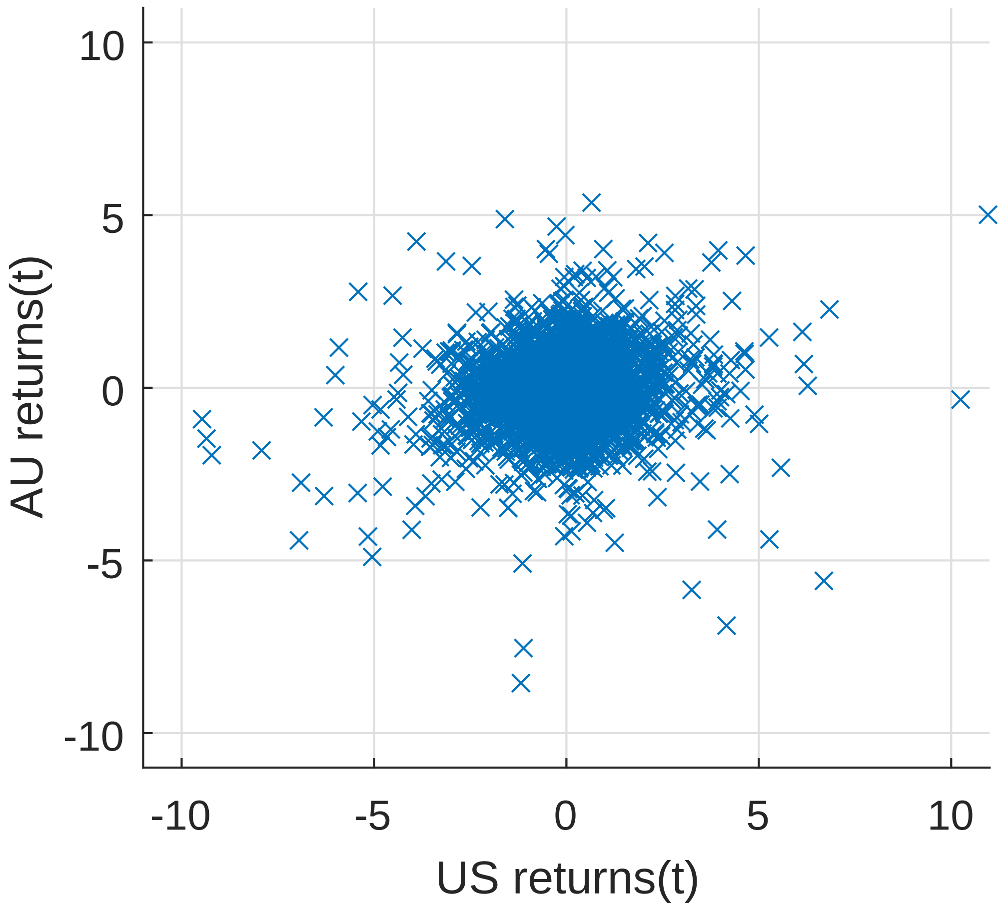

Figure 4: Scatterplot of contemporaneous daily log returns for AORD and SP500 indices

- The sample correlation between these series is 0.137, which is strongly significantly different to 0 (p-value = 0.000) at the 5% significance level.

- Figure 5 shows a scatterplot of AORD on day t vs **lagged** S&P500 daily log-returns (i.e. on day $t-1$), when the markets traded on these days.

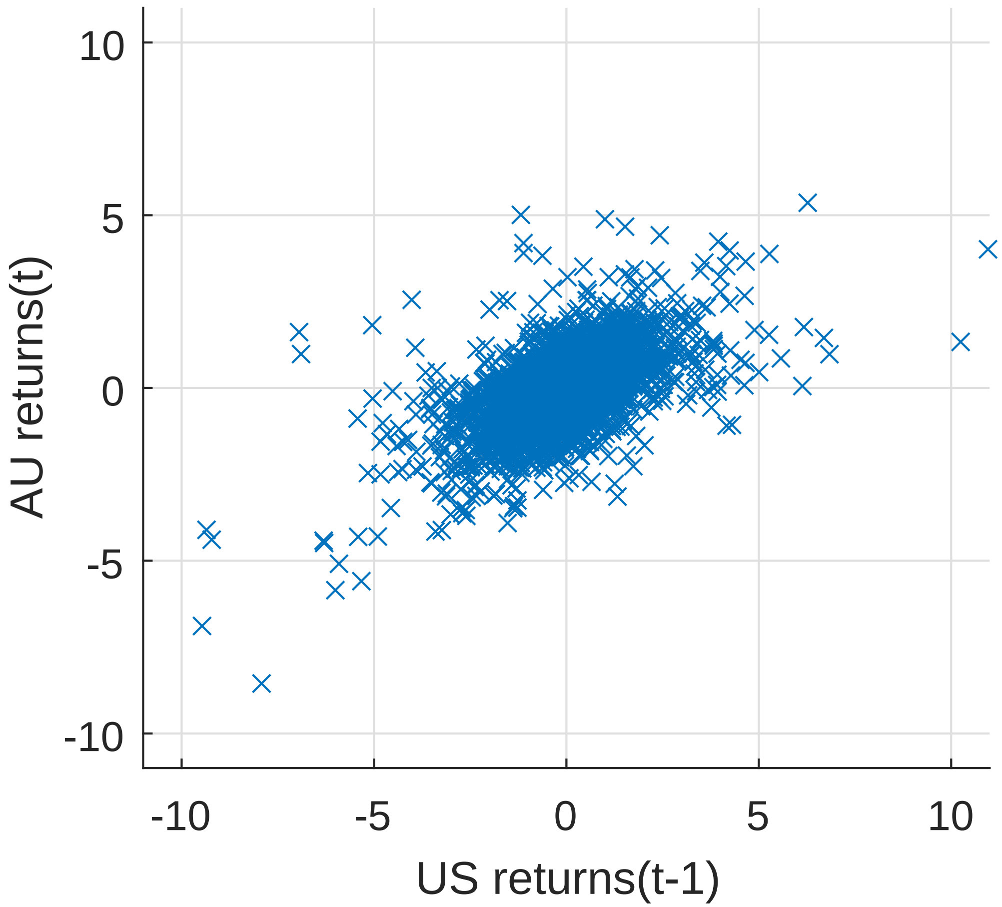

Figure 5: Scatterplot of daily log returns for AORD vs lagged SP500 indices

- The sample correlation is now 0.583, which is *strongly* and *practically* significantly different to 0 (p-value = 0.000).

- The plots, correlation values and tests indicate that ...?

- Example (ctd)

- In a SLR here, which should be the dependent and which the explanatory variable?

- The estimated regression relationship is:

$$\text{AORD}_t = 0.012 + 0.463 \times \text{S\&P500}_{t-1} + \widehat{\epsilon_t}$$

OR

$$\widehat{\text{AORD}}_t = 0.012 + 0.463 \times \text{S\&P500}_{t-1}$$

- A 1% increase in the return on S&P500 the day before, leads to an increase of 0.46% in the average AORD return on the next day.

- What about an increase of 5% on the S&P500 ??

- If the return on the S&P500 is 0%, the predicted return on the AORD the next day is 0.012%.

- Is this approach the only way to estimate the relationship here? Is it the best way? Issues? Assumptions?

- Is the relationship significant? Practically? Statistically?

- What is the strength of the relationship?

- Are the S&P500 daily price movements *causing* subsequent daily price movements in the AORD index? Or, is there another possible explanation?

Estimation: Least squares

- The most popular estimation method in regression is ordinary least squares (OLS)

- For the model:
$$y_t = \alpha + \beta X_t + \epsilon_t$$
estimates from a sample of data pairs $(X_1, y_1), \ldots, (X_T, y_T)$ are chosen that minimise:
$$\sum_{t=1}^{T} (y_t - \alpha - \beta X_t)^2$$

- These are often denoted $\hat{\alpha}$ and $\hat{\beta}$ and are called the OLS estimates

- To minimise the sum of squares is straightforward. Differentiate the expression in terms of α and then β, then set the two derivative equations to 0.

- Doing this results in the "normal" equations:

$$\sum_{t=1}^{T}(y_t - \alpha - \beta X_t) = 0$$

$$\sum_{t=1}^{T}(y_t - \alpha - \beta X_t)X_t = 0$$

and the resulting solutions:

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(y_t - \bar{y})(X_t - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{X}$$

Comments?

- Note that the sum of the estimated residuals is set to 0 in this case. Implications?

Some properties of OLS estimates

- Why are they popular?

- If the true correlation between X_t and the errors ϵ_t equals 0, then the OLS estimates are unbiased: i.e. $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$ when repeated over many samples.

- If the sample of data is iid and both variables X and y have finite 4th moments, then there is a central limit theorem for both $\hat{\alpha}$ and $\hat{\beta}$ that can be used for inference, t-tests and confidence intervals.

- i.e. Under the three LS assumptions (given below) we have $\hat{\alpha} \approx N(\alpha, \mathrm{Var}(\hat{\alpha}))$ and $\hat{\beta} \approx N(\beta, \mathrm{Var}(\hat{\beta}))$.

- Under these three assumptions, both $\hat{\alpha}$ and $\hat{\beta}$ tend to their respective true values

α and β, in very large samples.

- The three LS assumptions commonly applied in regression analyses:
  1. The independent variable X_t and the errors ϵ_t are uncorrelated, i.e. $E(\epsilon_t|X_t) = 0$
  2. The sample of data pairs $(X_1, y_1), \ldots, (X_T, y_T)$ is iid
  3. Both X and y have finite 4th moments, i.e. $E(X^4) < \infty$, $E(Y^4) < \infty$

- **These assumptions need to be acknowledged, discussed and assessed in each regression analysis you perform**

- These assumptions are automatically made by MATLAB (and most other software) when running a SLR analysis.

- Assumption 1 could only be the case if the error term ϵ_t contained factors that were uncorrelated with X_t

- Assumption 2 is satisfied by simple random sampling (SRS).

- Assumption 3 comes into doubt if the data are subject to extreme outliers.

- Assumption 3 implies that $E(Y)$, $E(Y^2)$, $\mathrm{Var}(Y)$ and $E(Y^3)$ are also all finite (as they are for X also).

- Figure 2 shows a seemingly innocuous pdf that has an infinite variance, i.e. for this rv Y represented: $E(Y^2) = \infty$ and $\mathrm{Var}(Y) = \infty$. Thus in this case also $E(Y^4) = \infty$.

- MATLAB also makes the assumption that $\mathrm{Var}(y_t|X_t) = \sigma^2$ is a constant when reporting inference for regression parameters.

- Example (ctd)

- The estimated regression relationship is:

$$\widehat{\mathrm{AORD}}_t = 0.012 + 0.463\mathrm{S\&P500}_{t-1}$$

- Is there a significant relationship?

- If the three LS assumptions hold, a 95% confidence interval for $\hat{\beta}$ is $(0.441, 0.486)$

- This indicates that we can be more than 95% confident that the true slope is greater than 0, and is instead somewhere in the range $(0.441, 0.486)$.

- YES, there is a statistically significant relationship: AORD returns increase significantly and proportionally with S&P500 returns, on average.

- The t-statistic for $\hat{\beta}$ is 40.2, indicating that the value 0.463 is 40.2 standard errors away from 0. The chance of seeing an estimated slope at least that far away from a true value of $\beta = 0$ is p-value $= 0$.

- The hypothesis that $\beta = 0$ has no support in the data and can be very strongly rejected.

- How strong is the relationship? R^2, SER

- $R^2 = 34\%$ indicating that 34% of the variation in AORD returns is captured by the straight line relationship with S&P500 returns.

- In other words, 34% of the daily movements in AORD index are explained by the previous day movements in S&P500 index.

- Note that

$$R^2 = 1 - \frac{\sum_{t=1}^{T}(y_t - \alpha - \beta X_t)^2}{\sum_{t=1}^{T}(y_t - \bar{y})^2} = \frac{\hat{Var}(y_t) - \hat{Var}(y_t|X_t)}{\hat{Var}(y_t)}$$

- R^2 directly measures how much $Var(y_t|X_t)$ is reduced from $Var(y_t)$ as a proportion. Here the reduction is by 34%.

Standard Error of Regrssion

- The $SER = 0.78\%$, meaning that the errors in predicting AORD returns, using the SLR model, have a standard deviation of 0.78%.

- SER is an estimate of the standard deviation of the residuals from the regression model.

残差

- Is the model a **strong** fit? Criteria?

- Does the model fit the data reasonably well? Residual plots

- Figure 6 shows the estimated errors (residuals) against the S&P500 returns.

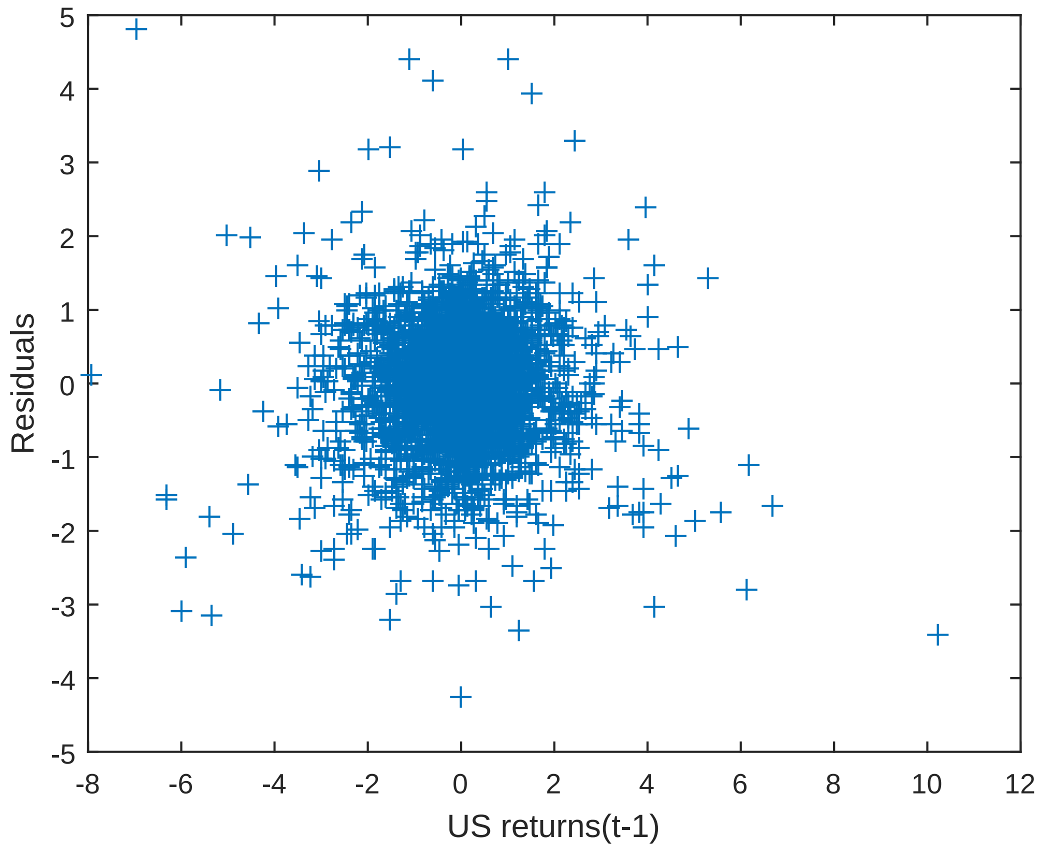

Figure 6: Scatterplot of residuals from SLR of daily log returns for AORD vs lagged SP500 indices

- A good fitting model would have $E(\hat{\epsilon}_t|X_t) = 0$ and hence show no pattern between the residuals and independent variable.

- Is this the case here?

- The residuals are formed by subtracting the fitted line value from each AORD return that it estimates, i.e. see Figure 7.

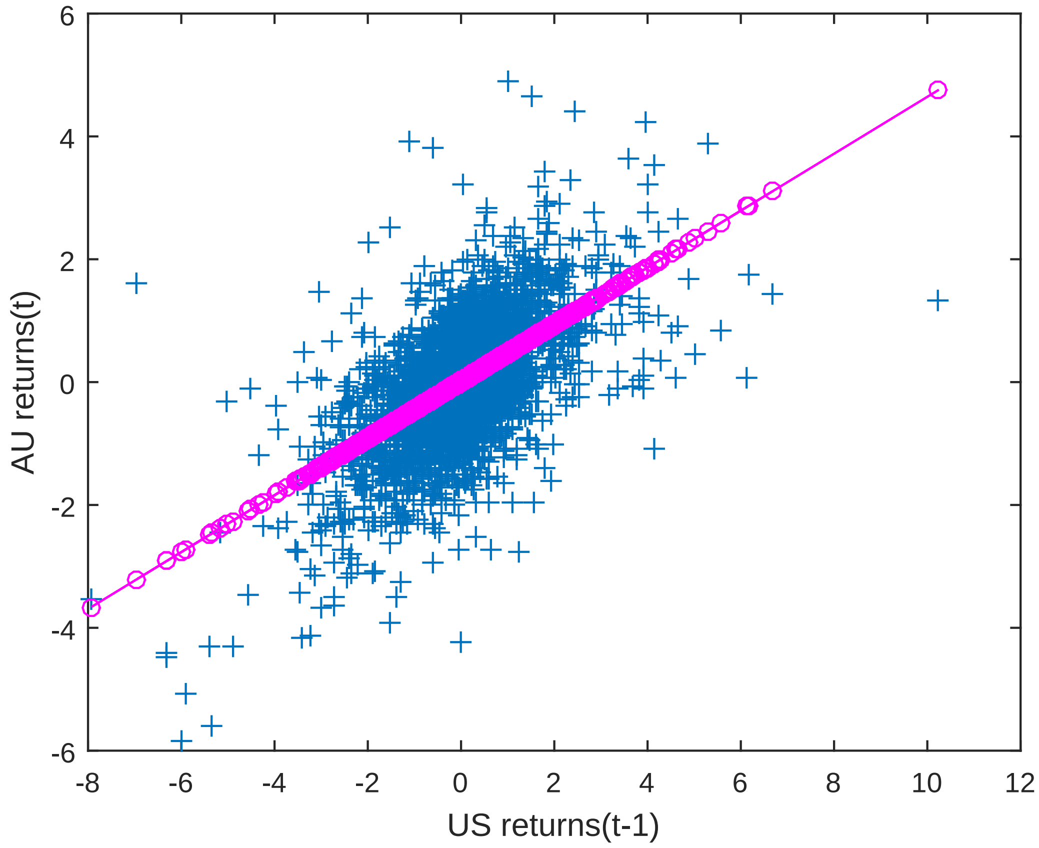

Figure 7: Scatterplot of daily log returns for AORD vs lagged SP500 indices plus SLR estimates

- Could the model be causal?

- Are there potential omitted variables that could be related to both AORD and S&P500 price movements?

2 Capital Asset Pricing Model

- The CAPM, or market model, is fundamental to modern asset pricing.
- The model relates excess asset returns to excess market returns, in an SLR.
- Let R_t^f be the risk-free rate of return and R_t^m be the market return at time t.
- An asset's excess return at time t is simply $R_t - R_t^f$.
- The CAPM is then written:

$$R_t - R_t^f = \alpha + \beta(R_t^m - R_t^f) + \epsilon_t$$

- OR, if $y_t = R_t - R_t^f$ and $X_t = R_t^m - R_t^f$, it is simply: $y_t = \alpha + \beta X_t + \epsilon_t$
- Here β is called the **market beta** and measures the sensitivity of the asset to

market movements.

- α is sometimes called Jensen's alpha and can be taken as a measure of above-market average profit.

- In particular, the unconditional average excess asset return is α plus β times the average excess market return.

- It is thus usual for financial analysts to look at both parameters in the CAPM,

- and perform tests of whether $\beta = 1$ and $\alpha = 0$

- $\beta > 1$ indicates high market risk, since market movements are amplified.

- $\alpha > 0$ is clearly preferred too.

Example

- Kenneth French's data library, is an excellent resource for US return data with proxies for market returns and risk-free rates of return.

- Consider the 5 industry sector asset portfolios in the US: Consumer, Manufacturing, HiTech, Health and Other.

- We consider each in the CAPM framework using daily data.

- The risk free rate proxy is the 1 month Treasury Bill rate, scaled to be a daily rate.

- First consider the Consumer sector portfolio:

- The scatterplot is shown in Figure 8, illustrating that ...

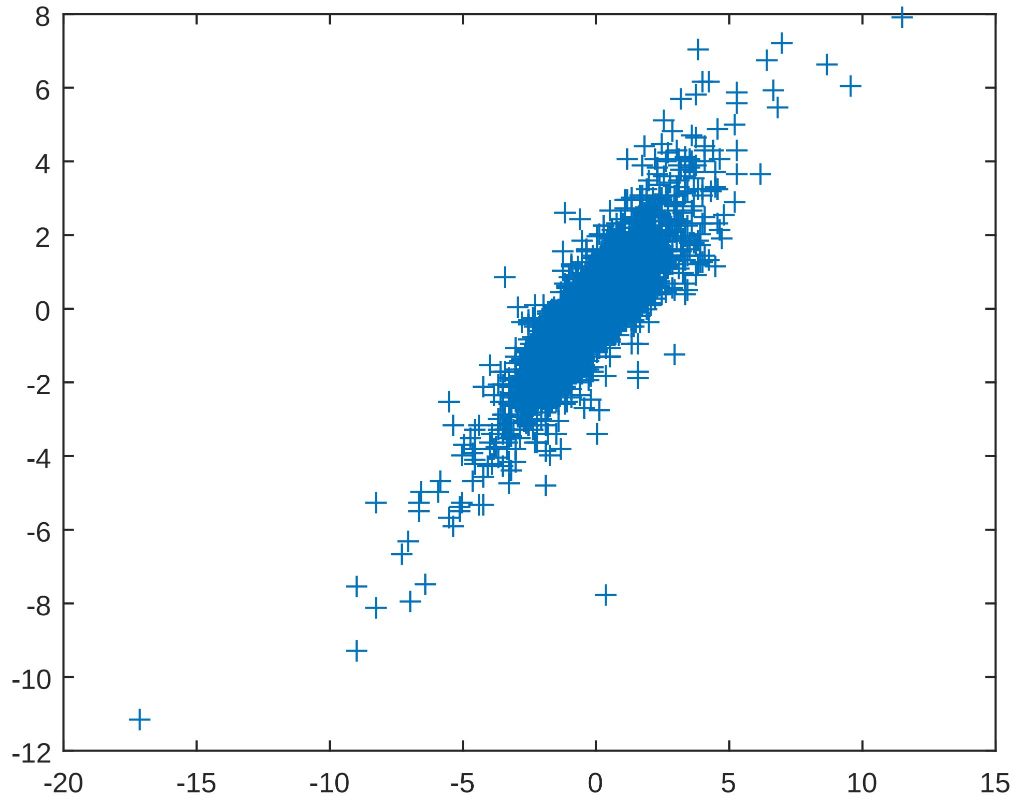

Figure 8: Scatterplot of daily excess returns for Consumer sector vs Market return

- The correlation between Consumer excess and Market excess returns is 0.85 (with p-value of 0)

- The estimated OLS regression/CAPM model is $\hat{y}_t = 0.04 + 0.75X_t$

- where y_t is excess consumer return and X_t is excess market return, both on day t.

- The market beta is 0.75, indicating that the consumer portfolio is not high risk. How sure are we of that?

- The 95% confidence interval for the true consumer market beta is $(0.741, 0.758)$. Thus we are more than 95% confident that the true market beta is less than 1.

- It is highly likely that consumer portfolio is less volatile or risky than the market portfolio, on average.

- Further, the estimated excess return when the excess market return is 0 is $\hat{\alpha} = 0.04\%$.

- The 95% CI for α is $(0.031, 0.048)$: we are at least 95% certain that the average Consumer portfolio excess return is generally higher than the average market excess return (certainly when the latter is close to the risk free rate, right?).

- But not by much!

- Figures 9 and 10 respectively show the estimated regression line and the residual plot.

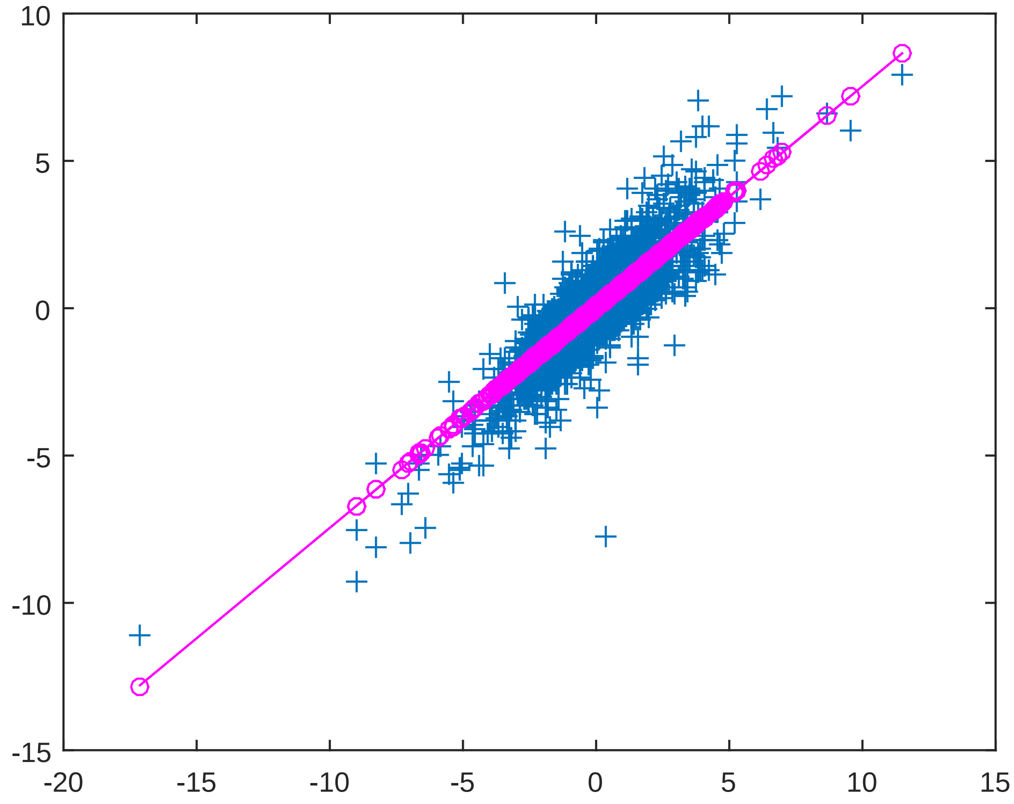

Figure 9: Scatterplot of daily excess returns for Consumer sector vs Market return plus OLS regression line

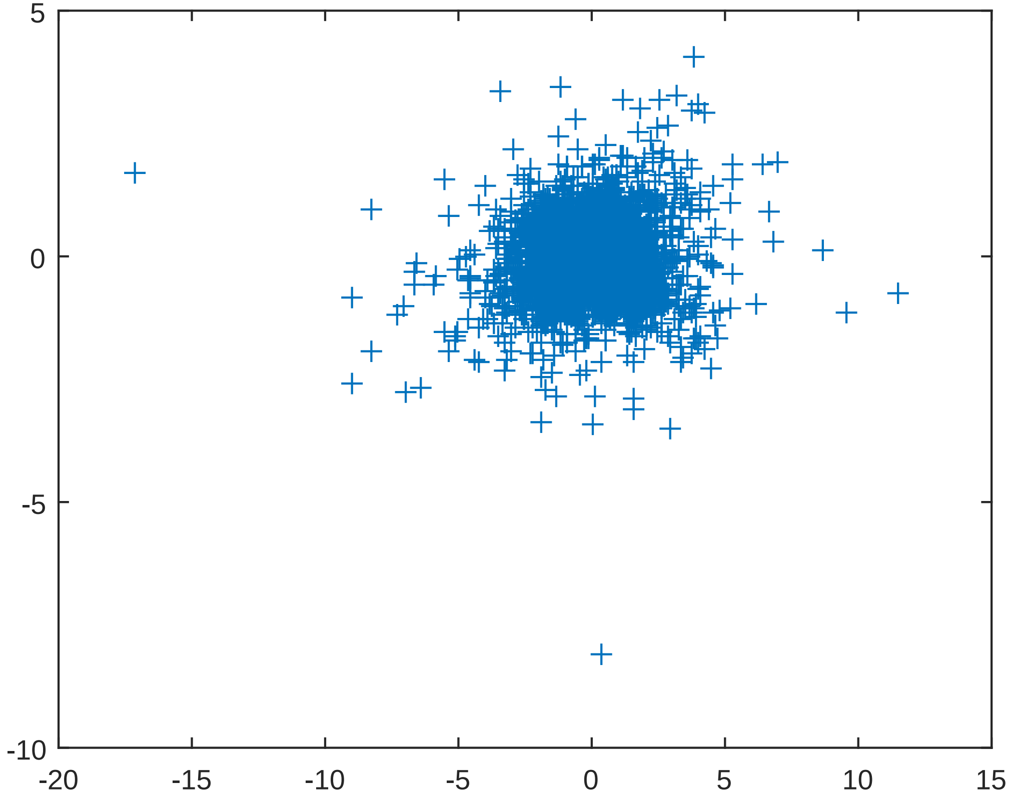

Figure 10: Scatterplot of residuals vs Market return for Consumer CAPM

- Does the model fit the data well?

Table 1: CAPM estimates for 5 daily industry portfolios

Industry	α	CI for α	β	CI for β	R^2	SER
Consumer	0.040	(0.031,0.048)	0.749	(0.741,0.758)	0.72	0.46
Manufacturing	0.040	(0.031,0.049)	0.812	(0.804,0.821)	0.73	0.49
Hi-Tech	0.053	(0.042,0.065)	0.988	(0.976,0.9996)	0.69	0.66
Health	0.054	(0.043,0.065)	0.833	(0.822,0.844)	0.64	0.63
Other	0.046	(0.038,0.054)	0.699	(0.691,0.707)	0.70	0.46

- The $R^2 = 72\%$ and the $SER = 0.46\%$. Is the model a strong fit?

- Table 2 shows Jensen's alpha and market beta for each of the 5 industry portfolios.

- All the portfolios have Jensen's alpha significantly greater than 0%.

- All have market betas significantly less than 1 (Hi-Tech only just).

- All have market beta significantly different to 0.

- Is HiTech a low risk portfolio? Does it have market beta < 1? We can use a t-test here also

- Alternative hypothesis is that $\beta < 1$, null is that $\beta = 1$.

- The t-statistic is given by:

$$\begin{aligned} t &= \frac{\hat{\beta} - 1}{\sqrt{\text{Var}(\hat{\beta})}} = \frac{\hat{\beta} - 1}{\text{SE}(\hat{\beta})} \\ &= \frac{0.988 - 1}{0.006} \end{aligned}$$

which is $t = -2.034$.

- The p-value is approximately the probability of observing a random Gaussian lower than -2.034, which here is 0.021.

- Thus, at the 5% significance level, we can conclude that the market beta for HiTech is significantly less than 1. It is very likely to be a low risk or defensive portfolio.

- Can we trust these results?

- Are the sets of asset return pairs iid?

- Are there omitted factors that are, or might be, also correlated with the market return?

- Are there large outliers that may cast doubt that up to 4th moments of these returns are finite?

- What should we do if there are outliers? Remove them? Something else?

- In finance, outliers should ONLY be removed if you do not want your investment strategy, risk measurements, etc to be accurate or valid **in the cases of those outlying returns**.

- e.g.: DON'T remove financial crises; Possibly remove stock-splits; Definitely remove known data errors.

- If outliers are an issue, robust methods, e.g. quantile regression, can be employed instead.

- Outside this unit scope but see 'quantilereg_RK.m'. Also 'robust' in MATLAB's Statistics toolbox.

3 Multi-factor models

- Simple regression can easily be extended to multiple explanatory factors.

- Called multiple regression:

$$y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \epsilon_t,$$

  is a 2 factor model.

- OR in general

$$y_t = \alpha + \beta_1 X_{1,t} + \beta_2 X_{2,t} + \ldots + \beta_m X_{m,t} + \epsilon_t,$$

  which is an m-factor model.

- The most commonly applied estimator is again OLS.

- OLS estimates from a sample of T data points

$$(X_{1,1}, \ldots, X_{m,1}, y_1), \ldots, (X_{1,T}, \ldots, X_{m,T}, y_T)$$

  are chosen to minimise:

$$\sum_{t=1}^{T}(y_t - \alpha - \beta_1 X_{1,t} - \beta_2 X_{2,t} - \ldots - \beta_m X_{m,t})^2$$

- These OLS estimates are often denoted $\hat{\alpha}$ and $\hat{\beta}_1, \ldots, \hat{\beta}_m$

- OR in vector form, $\hat{\boldsymbol{\beta}}$

- To minimise the sum of squares equation is achieved by differentiating the expression separately in terms of each element of β, then setting these $m + 1$ derivatives equations to 0.

- This can be done exactly and uniquely, as long as ....

- In matrix form we have:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \; ; \; \boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & \dots & X_{m,1} \\ 1 & X_{1,2} & X_{2,2} & \dots & X_{m,2} \\ \vdots & \vdots & \dots & \vdots & \\ 1 & X_{1,T} & X_{2,T} & \dots & X_{m,T} \end{pmatrix}$$

- The model can then be written more efficiently as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- Then, differentiating the sum of squares equation and setting these derivatives to 0 leads to the "normal" equations:

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}$$

- Using the rules of matrix algebra, the unique solution is:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

- This unique solution exists if, and only if, the matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists.

- This matrix inversion exists if and only if, roughly, each variable $\boldsymbol{X}_i$ contributes information that is not already provided by the other variables in $\boldsymbol{X}$.

- This means that none of the variables in, i.e. the columns of, $\boldsymbol{X}$ can be perfectly correlated with any other.

- Also none can be the exact linear combination of any of the other variables in $\boldsymbol{X}$. Also, none can be fixed at a constant value (except the first column).

- These last points make sense because the interpretation of β_i is the effect on y of changing X_i by one unit, **holding all other variables in $\boldsymbol{X}$ constant**.

- This would not be possible under the conditions of perfect correlation, linear combinations or constancy.

- Strictly speaking, this means the columns in $\boldsymbol{X}$ must be linearly independent of each other.

- This becomes the fourth assumption required to do OLS estimation and inference. The other three remain the same.

- For some background on matrices and vectors, see the matrix notes in the Supplementary section on Canvas.

Example

- Again we employ Kenneth French's data library and consider the 5 industry sector asset portfolios in the US: Consumer, Manufacturing, HiTech, Health and Other.

- We consider each in the multi-factor CAPM framework.

- Fama and French (1992, 1993) consider factors that they thought might be influencing asset returns.

- These include: 关于市值的小型和大型股票投资组合收益率之间的差异

  1. SMB: the difference between returns on portfolios of small and big stocks, regarding market capitalization
  2. HML: The difference between returns on portfolios of high and low stocks, re book-market ratios

     高股票和低股票投资组合的回报，再保险市场比率之间的差异

  among others.

- We add these to the single factor CAPMs we considered above.

- The scatterplot is shown in Figure 11, showing ...

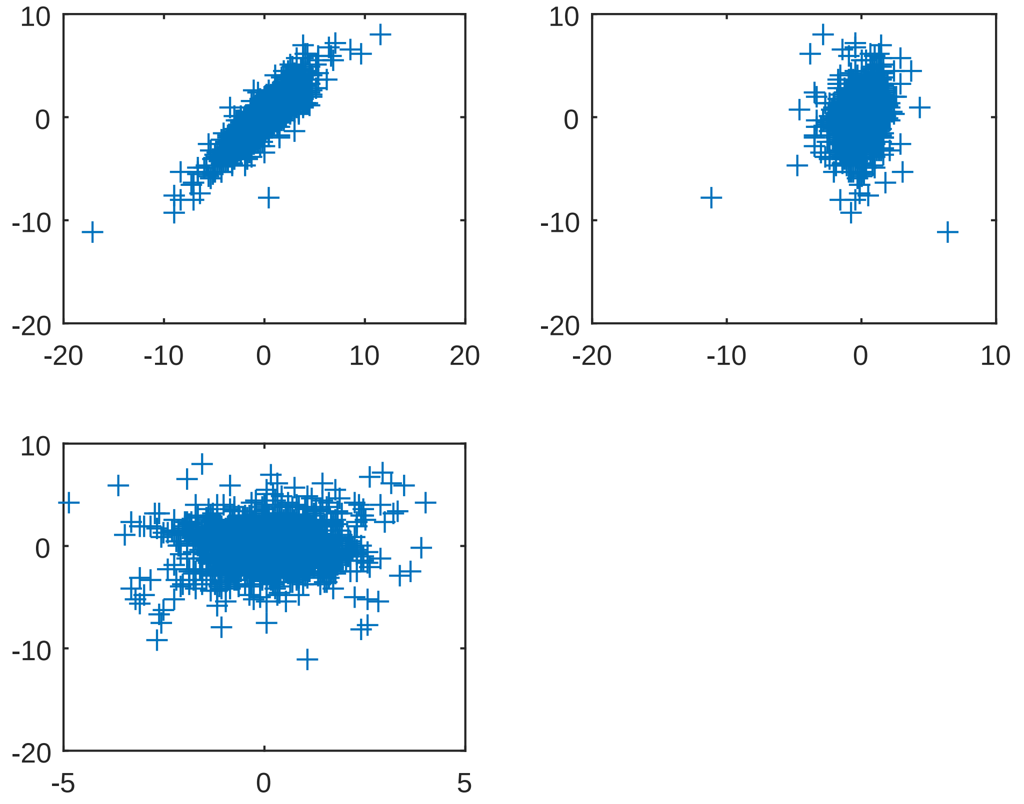

Figure 11: Scatterplot of daily excess returns for Consumer sector vs Market return; vs HML and vs SMB

- The correlation between Consumer excess returns and market excess returns is 0.85 (p-value = 0); between SMB and Consumer excess returns is 0.25 (p-val=0) and between Consumer and HML is -0.12 (p-val=0)

- The estimated OLS regression/CAPM model is

$$\hat{y}_t = 0.03 + 0.85(R_t^m - R_t^f) + 0.69 smb_t + 0.32 hml_t$$

- where y_t is excess consumer return, $R_t^m - R_t^f$ is excess market return and smb_t and hml_t are the differences in returns from the small and large cap, and high and low book to market, portfolios, respectively.

- The market beta is now estimated as 0.85, indicating that the consumer portfolio is not high risk. How sure are we of that?

- The 95% confidence interval for the true market beta of consumer is $(0.841, 0.852)$.

Thus we are still at least 95% confident that the true market beta is less than 1.

- Further, the estimated excess return when the excess market return is 0 is $\hat{\alpha} = 0.027\%$.

- The 95% CI for α is $(0.022, 0.032)$ and thus we are at least 95% certain that the average Consumer portfolio excess return is higher than the average excess return on the market.

- Figure 12 shows the consumer excess returns, the original estimated single factor CAPM line, plus the new estimates from the multi-factor model.

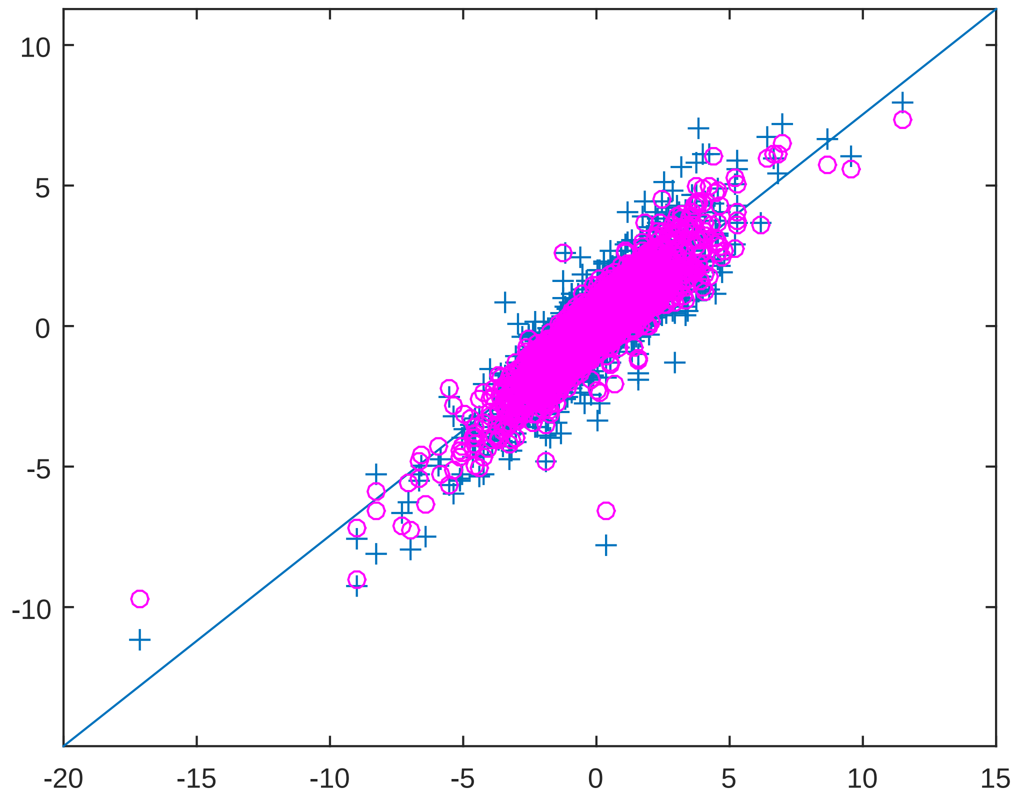

Figure 12: Scatterplot of daily excess returns for Consumer sector vs Market return plus OLS regression estimates

- Figure 13 shows the residual plot from the new model.

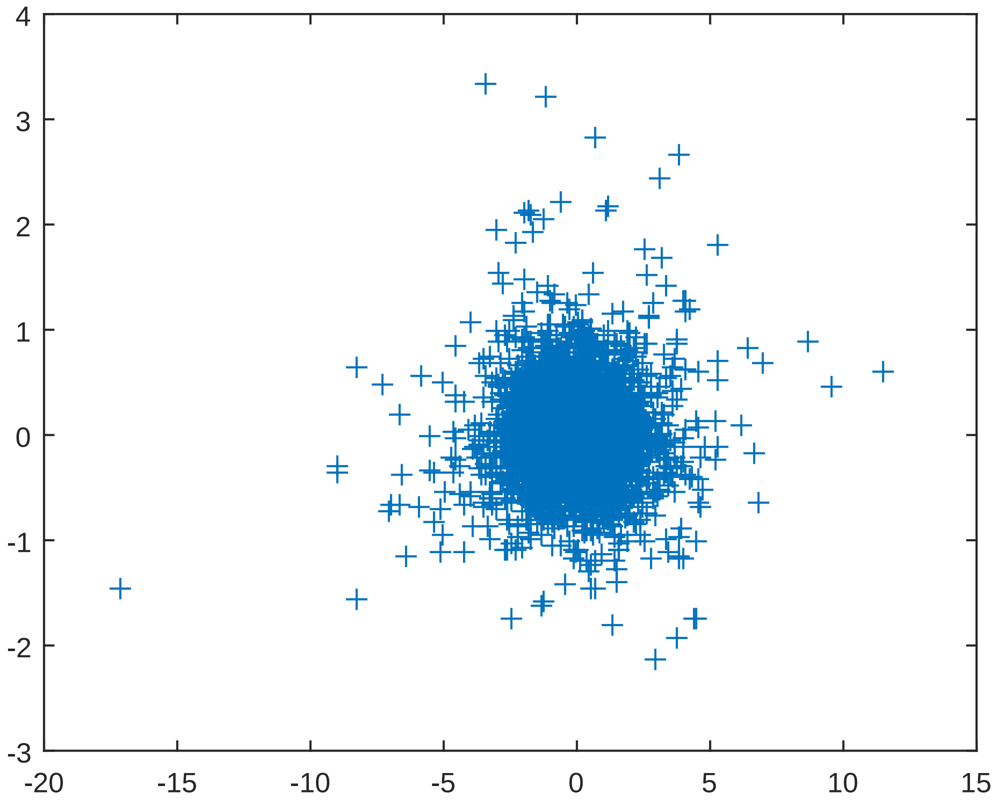

Figure 13: Scatterplot of residuals vs Market return for Consumer three factor CAPM

- Does the model fit the data well? Has it improved the fit over the simple single-factor CAPM?

- The $R^2 = 90\%$ and the $SER = 0.41\%$. Is the model a strong fit?

- Table 2 shows parameter estimates for each of the 5 industry portfolios from the multi-factor CAPM.

Table 2: CAPM estimates for 5 daily industry portfolios

Industry	$\hat{\alpha}$ CI for α	$\hat{\beta}_1$ CI for β_1	$\hat{\beta}_2$ CI for β_2	$\hat{\beta}_3$ CI for β_3	R^2 (old)	SER (old)	R^2 (adj)
Consumer	0.027	0.846	0.691	0.315	0.89	0.29	0.89
	(0.02,0.03)	(0.84,0.85)	(0.68, 0.70)	(0.30, 0.33)	(0.72)	(0.46)	
Manufacturing	0.025	0.923	0.633	0.440	0.87	0.33	0.87
	(0.02,0.03)	(0.92,0.93)	(0.62, 0.65)	(0.43, 0.45)	(0.73)	(0.49)	
Hi-Tech	0.049	1.031	0.956	-0.203	0.87	0.42	0.87
	(0.04,0.06)	(1.02,1.04)	(0.94, 0.97)	(-0.22, -0.19)	(0.69)	(0.66)	
Health	0.050	0.874	0.764	-0.120	0.78	0.49	0.78
	(0.04,0.06)	(0.86,0.88)	(0.75, 0.78)	(-0.14, -0.10)	(0.64)	(0.63)	
Other	0.033	0.799	0.652	0.359	0.88	0.29	0.88
	(0.03,0.04)	(0.79,0.80)	(0.64, 0.66)	(0.35, 0.37)	(0.70)	(0.46)	

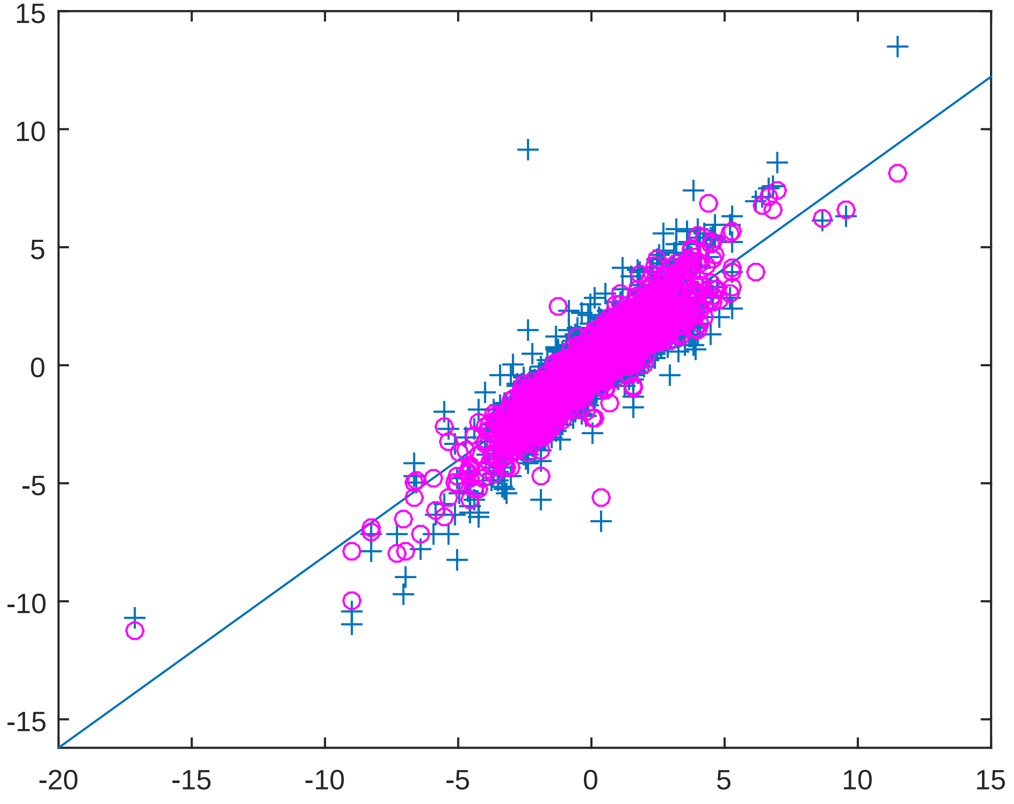

Figure 14: Scatterplot of daily excess returns for Manufacturing sector vs Market return plus OLS regression estimates

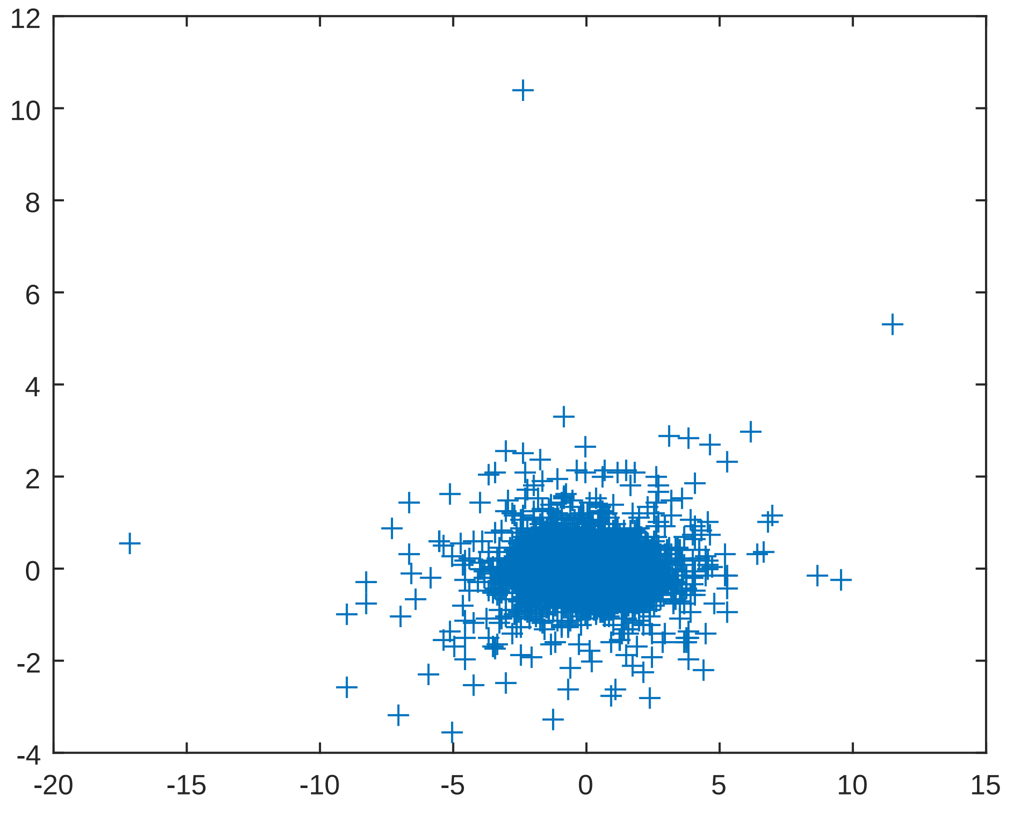

Figure 15: Scatterplot of residuals vs Market return for Manufacturing three factor CAPM

- All the portfolios still have Jensen's alpha significantly greater than 0%, though the interpretation is now the average excess return when the market excess return is 0% AND the differences in return on HML and on SMB are also 0%.

- All portfolios now have market beta significantly less than 1, except for HiTech which is significantly greater than 1.

- All have market beta significantly different to 0.

- All have coefficients on SMB and HML significantly different to 0.

- All portfolios have multi-factor CAPMs fitting more strongly than single factor CAPM by R^2 and SER

- R^2 ALWAYS increases when extra factors are added and then fit to the SAME data.

- Adjusted R^2 is often used instead. This always increases when SER decreases, when two models are applied on the same data.

- Like SER, Adjusted R^2 can either increase or decrease when an extra regressor is added to the model.

- The formula for this is:

$$\text{Adjusted}R^2 = 1 - \frac{SER^2}{s_Y^2} = \frac{\widehat{Var(y_t)} - \widehat{Var(y_t|X_t)}}{\widehat{Var(y_t)}}$$

- The interpretation is **exactly** the same as that for R^2. This just uses a different, and in fact unbiased, estimator of $Var(y_t|X_t)$ which is

$$\widehat{Var(y_t|X_t)} = SER^2 = \frac{\sum_{t=1}^{T}(y_t - \alpha - \beta_1 X_{1,t} - \ldots - \beta_m X_{m,t})^2}{T - m - 1}$$

.

- For large sample sizes, $R^2 \approx R^2$-adjusted, as reflected in the table above

- Is HiTech a high risk portfolio? Does it have market beta > 1?

- Yes, no need to test, since we can see that the 95% confidence interval does not include 1 (both upper and lower limits < 1) so the data do not support the case of beta > 1.

- Can we trust these results?

- We could trust these results if and only if the three LS assumptions held in each case.

- This means there are no omitted variables that are also correlated with ANY of the market, HML or SMB return series

- And that the returns in all series are iid. What does financial theory say about that?

- And that the fourth moments are finite in each series, i.e. outliers are rare.

- We'll look at these questions in more detail in lab.

- Can we use these results to make investment strategies? Do risk management? Something else beneficial?

4 Stress testing

- Stress testing means the process of subjecting a model to very unlikely extreme events, so as to learn how the asset or portfolio or index being modelled might react.

- It can be applied to very simple models, like the CAPM, but is more usually applied to large complex models that include macro-economic factors, etc. These are beyond this unit.

- As a simple example, consider the ordinary CAPM:

$$R_t - R_t^f = \alpha + \beta(R_t^m - R_t^f) + \epsilon_t$$

- OR, if $y_t = R_t - R_t^f$ and $X_t = R_t^m - R_t^f$, it is simply: $y_t = \alpha + \beta X_t + \epsilon_t$

- What might happen to the excess return $R_t - R_t^f$ if the (excess) market return was -10% in one day?

- The CAPM predicts:

$$E(R_t - R_t^f | R_t^m - R_t^f = -10) = \alpha - 10 \times \beta$$

  i.e. the expected excess return when the market drops by 10% (below risk-free rate) is simply $\alpha - 10 \times \beta$

- What about other quantities, like $\text{Var}(y_t | X_t = -10)$? What is the lowest return we expect to occur, say 1 out of 100 days, when the market drops by 10% in one day? Or, $E(y_t|X_t = -10 \text{AND} y_t < Q_y(0.01))$

- The 2nd quantity is called Value at Risk (VaR)

- The 3rd quantity is the expected or average loss when the market drops by 10% AND the asset return is below its first quantile VaR. It is called expected shortfall (ES), or conditional VaR.

- These three quantities are the most common quantitative measures of risk.

- The CAPM model as we estimate it in MATLAB suggests that $\text{Var}(y_t|X_t = -10) = \sigma^2$. This implies ...?

- We'll see how to improve this assumption later in the unit.

- So, far the CAPM says nothing about VaR or ES, as we have specified it.

- There are two approaches we could use to make it do so:
  1. Parametric: assume a specific distribution for $y_t|X_t$ or ϵ_t (same thing).
  2. Non-parametric: use properties of the data to estimate VaR, ES without assuming a specific distribution

- One parametric choice is the Gaussian distribution, i.e. assume that $\epsilon_t \sim N(0, \sigma^2)$

- In that case,
$$\text{VaR}_t(0.01) = \alpha - 10 \times \beta - 2.326\sigma$$
since $\Phi^{-1}(0.01) = -2.326$

- This is because the 1% quantile of a Gaussian distribution, $N(\mu, \sigma^2)$, lies at

$\mu - 2.326\sigma$

- Also, the ES at 1% level is

$$\text{ES}_t(0.01) = \alpha - 10 \times \beta - 2.667\sigma$$

- This is because the expected value below the 1% quantile of a Gaussian distribution, lies at the 0.38% quantile of the $N(\mu, \sigma^2)$, and hence $\text{ES}(0.01) = \mu - 2.667\sigma$, since $\Phi^{-1}(0.0038) = -2.667$

- For all distributions we must have $\text{ES}(\alpha) \leq \text{VaR}(\alpha)$, by definition.

- Figure 16 illustrates the concepts of VaR and ES

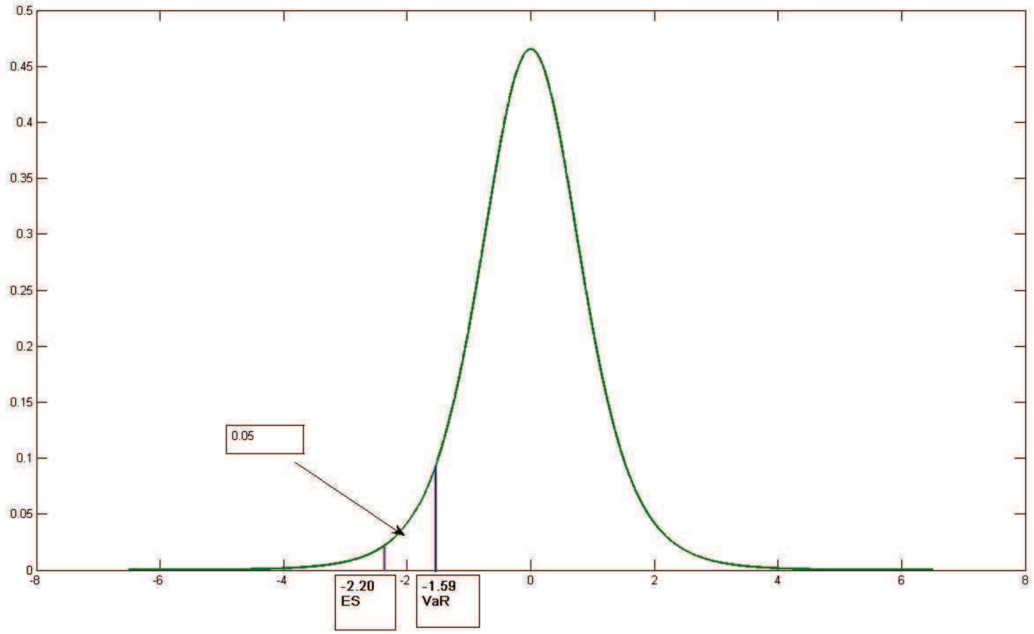

Figure 16: Illustration of VaR and ES at 5% level

- A non-parametric approach could be as follows:

- Use the 1st quantile of the residual series $\hat{\epsilon}_t$ to estimate: $\text{VaR}_t(0.01) = \alpha - 10 \times \beta - Q_{\hat{\epsilon}}(0.01)$

- Further, estimate the sample mean residual below $Q_{\hat{\epsilon}}(0.01)$ and utilise in: $\text{ES}_t(0.01) = \alpha - 10 \times \beta - \bar{\hat{\epsilon}}(\hat{\epsilon} < Q_{\hat{\epsilon}}(0.01))$

- Single factor CAPM example:

- Table 3 again shows the estimated CAPM for each of the 5 industry portfolios.

Table 3: CAPM estimates for 5 daily industry portfolios

Industry	α	CI for α	β	CI for β	R^2	SER
Consumer	0.040	(0.031,0.048)	0.749	(0.741,0.758)	0.72	0.46
Manufacturing	0.040	(0.031,0.049)	0.812	(0.804,0.821)	0.73	0.49
Hi-Tech	0.053	(0.042,0.065)	0.988	(0.976,0.9996)	0.69	0.66
Health	0.054	(0.043,0.065)	0.833	(0.822,0.844)	0.64	0.63
Other	0.046	(0.038,0.054)	0.699	(0.691,0.707)	0.70	0.46

- Based on these, and assuming a conditional Gaussian distribution for each industry portfolio, and a market excess return of -10%, we have the estimated stress testing quantities, in Table 4

Table 4: CAPM risk estimates for 5 daily industry portfolios under Gaussian distribution and -10% market excess return

			Gaussian		Non-parametric	
Industry	E(excess return)	σ^2	VaR (0.01)	ES(0.01)	VaR (0.01)	ES(0.01)
Consumer	-7.45	0.22	-8.54	-8.69	-8.71	-9.25
Manufacturing	-8.08	0.24	-9.23	-9.40	-9.49	-10.02
Hi-Tech	-9.83	0.43	-11.36	-11.59	-11.62	-12.42
Health	-8.27	0.39	-9.73	-9.95	-9.92	-10.63
Other	-6.94	0.21	-8.00	-8.16	-8.14	-8.67

- By all measures, the Hi-tech industry portfolio seems to be the most at risk to a large single day drop in the market.

- Also, the Other portfolio seems least exposed to large single day drops in the market index.

- Can you/your company survive such a drop in the market index? Can you hedge this risk somehow?

- We will consider other approaches to estimate risk measures and stress test measures later in the unit.

- And we will examine VaR and ES in more detail and more deeply.

- Note that a quantile regression approach could have been taken instead of sample quantiles for the non-parametric approach above.