Factor models and linear regression

Lecture 8

Dr. Martin Ewers

April 23, 2014

Table of contents

1. Notation

2. Least squares Regression

2.1 Definition ‘regression’

2.2 Noise

2.3 Optimization calculus

3. Example: SIM

4. Multifactor regression

4.1 Model input

4.2 Optimization calculus

5. Quality of regression

5.1 Coefficient of determination

5.2 Correlation coefficient

Section 1

Notation

Notation

Random variables

r_0	riskless return
R_i	return asset i
R_M	return indexed portfolio (e.g. DAX)
$ER_i = R_i - r_0$	excess return asset i
$ER_M = R_M - r_0$	excess return indexed portfolio
ε_i	idiosyncratic component of ER_i

Notation

Observed values

$r_{0,t}$	observed riskless return
$R_{i,t}$	observed return asset i
$R_{M,t}$	observed return indexed portfolio
$ER_{i,t} = R_{i,t} - r_{0,t}$	observed excess return i
$ER_{M,t} = R_{M,t} - r_{0,t}$	observed excess return indexed portfolio
$\varepsilon_{i,t}$	residuals

Notation

$\bar{R}_i$	sample mean
$\text{var}(R_i) = \sum_{t=1}^{m} \frac{1}{m} \cdot (R_{i,t} - \bar{R}_i)^2$	sample variance of R_i
$\text{var}(F) = \sum_{t=1}^{m} \frac{1}{m} \cdot (F_t - \bar{F})^2$	sample variance of F
$\text{cov}(R_i, F) = \sum_{t=1}^{m} \frac{1}{m} \cdot (R_{i,t} - \bar{R}_i) \cdot (F_t - \bar{F})$	sample covariance of F and R_i

Notation

$s_i = \sqrt{\text{var}(R_i)}$	sample standard deviation of R_i
$s_F = \sqrt{\text{var}(F)}$	sample standard deviation of F

Section 2

Least squares Regression

Least squares Regression

Definition ‘regression’

Definition (Regression)

- Technique used for the modelling and analysis of numerical data
- Exploits the relationship between two or more variables so that we can gain information about one of them through knowing values of the other
- Regression can be used for prediction, estimation, hypothesis testing, and modelling causal relationships

Least squares Regression

Definition 'regression'

Regression line

$\hat{R}_i$ is a prognosis of the return for a given value of F:

$$\hat{R}_i = a_i + b_i \cdot F$$

Assumption: The sample distribution of F_t ($t = 0, \ldots, n$) corresponds to the probability distribution of the random variable F.

Conclusion: $\hat{R}_i = \mathrm{E}[R|F]$ (conditional expected value)

Least squares Regression

Noise

Explained versus unexplained sample variation

$$R_{i,t} = a_i + b_i \cdot F_t + \varepsilon_{i,t}$$

$\hat{R}_{i,t} - \bar{R}_i$: explained sample variation

$R_{i,t} - \hat{R}_i = \varepsilon_{i,t}$: unexplained sample variation (colloquial: 'noise')

$R_{i,t} - \bar{R}_i$: total sample variation

Least squares regression

Optimization calculus

Least squares regression means: The sum of the squared residuals $\varepsilon_{i,t}$ gets minimised.

$$SSR_i = \sum_{t=1}^{m} (\varepsilon_{i,t})^2 = \sum_{i=1}^{m} (R_{i,t} - \hat{R}_{i,t})^2 \rightarrow \min_{a_i, b_i}!$$

with

m number of observations in scatter plot

Least squares regression

Optimization calculus

Conditions for a minimum of SSR_i

$$\begin{cases} \dfrac{\partial SSR_i}{\partial a_i} & = 0 \\[2ex] \dfrac{\partial SSR_i}{\partial b_i} & = 0 \end{cases}$$

$$\Rightarrow \quad \begin{cases} a_i & = \bar{R}_i - b_i \cdot \bar{F} \\ b_i & = \dfrac{\mathrm{cov}(R_i, F)}{\mathrm{var}(F)} \end{cases}$$

Section 3

Example: SIM

Example: SIM

SIM – Basic assumption

$$ER_i = a_i + b_i \cdot ER_M + \varepsilon_i$$

with

$ER_i = R_i - r_0$	random excess return on asset i
$ER_M = R_M - r_0$	random excess return on indexed portfolio

Example: SIM

Definition (Security Characteristic Line)

The Security Characteristic Line is regression line

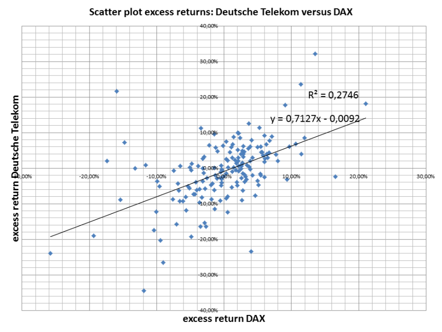
$$\widehat{ER}_i = a_i + b_i \cdot ER_M$$

CAPM: Security Market Line

Single Index Model: Security Characteristic Line

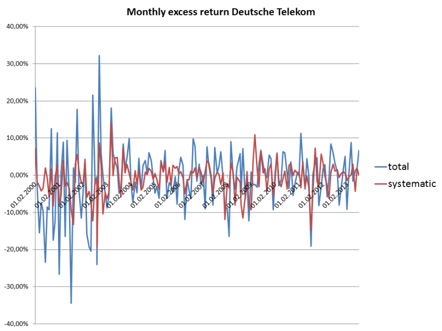
Example: SIM

Scatter plot $(ER_{M,t}, ER_{i,t})$



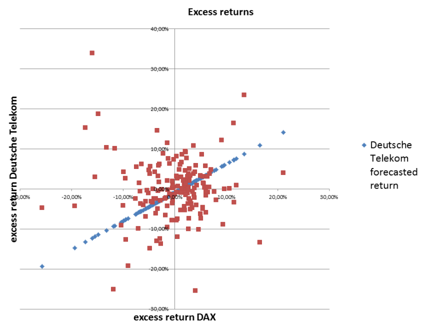
Example: SIM

$ER_{i,t}$ versus residuals $\varepsilon_{i,t}$



Example: SIM

The residuals $\varepsilon_{i,t}$ are not determined by $ER_{i,t}$



Example: SIM

Interpretation of b_i

General:

$$\frac{\partial \widehat{ER}_i}{\partial ER_M} = b_i$$

Example: $b_{DT} = 0.7127$, with DT denoting ‘Deutsche Telekom’.

- DT is less volatile than the DAX.
- For every 1% change in the DAX, we expect the return on DT to change by 0.7127%.

Example: SIM

Interpretation of a_i in case of a SIM

General:

$$ER_M = 0 \Rightarrow \widehat{ER}_i(0) = a_i$$

Example: $a_{DT} = -0.0092$.

- $ER_M = 0$ means the DAX does not change, which is a random event.
- If $ER_M = 0$, DT's return is expected to be -0.0092.
- Conclusion: Other factors than the DAX have a negative impact on DT's return.

Example: SIM

Estimating b_i

Recall that for random variables R_i and $ER_i = R_i - r_0$:

$$b_i = \frac{\mathrm{COV}(R_i, R_M)}{\mathrm{VAR}(R_M)} = \frac{\mathrm{COV}(ER_i, ER_M)}{\mathrm{VAR}(ER_M)}$$

Background: The riskless return r_0 is by definition no random variable.

Example: SIM

Estimating b_i

b_i might be estimated either on the basis of

- observed raw returns $R_{i,t}$ and $R_{M,t}$ or
- observed excess returns $ER_{i,t}$ and $ER_{M,t}$

These approaches will yield different results:

- Historically, the riskless rate takes different values.
- Correspondingly, $r_{0,t}$ will take different values in the sample.

Section 4

Multifactor regression

Method of least squares

Model input

n observations of m factors as model input

$(F_{1,t}, \ldots, F_{m,t}, R_{i,t})$ is a pair of observed values. Each of the following n equations corresponds to an observation $(t = 1, \ldots, n)$:

$$
\begin{array}{lllllll}
\varepsilon_{i,1} = & R_{i,1} & +a_{i,0} & +b_{i,1} \cdot F_{1,1} + & \ldots & +b_{i,m} \cdot F_{m,1} \\
\varepsilon_{i,2} = & R_{i,2} & +a_{i,0} & +b_{i,1} \cdot F_{1,2} + & \ldots & +b_{i,m} \cdot F_{m,2} \\
\vdots & \vdots & \vdots & \vdots & & \vdots \\
\varepsilon_{i,n} = & R_{i,n} & +a_{i,0} & +b_{i,1} \cdot F_{1,n} + & \ldots & +b_{i,m} \cdot F_{m,n}
\end{array}
$$

Method of least squares

Model input

n observations as model input

$$\vec{\varepsilon}_i = \vec{R}_i + M \cdot \vec{k}_i$$

specific to asset i	common to all assets in analysed portfolio
$\vec{\varepsilon}_i = \begin{pmatrix} \varepsilon_{i,1} \\ \varepsilon_{i,2} \\ \vdots \\ \varepsilon_{i,n} \end{pmatrix}; \vec{R}_i = \begin{pmatrix} R_{i,1} \\ R_{i,2} \\ \vdots \\ R_{i,n} \end{pmatrix}; \vec{k}_i = \begin{pmatrix} a_{i,0} \\ b_{i,1} \\ b_{i,2} \\ \vdots \\ b_{i,m} \end{pmatrix}$	$M = \begin{pmatrix} 1 & F_{1,1} & \dots & F_{m,1} \\ 1 & F_{1,2} & \dots & F_{m,2} \\ \vdots & \vdots & & \vdots \\ 1 & F_{1,n} & \dots & F_{m,n} \end{pmatrix}$

Method of least squares

Optimization calculus

Least squares regression means: The sum of the squared residuals $\varepsilon_{i,t}$ gets minimised.

$$SSR_i = f(\vec{k}_i) \to \min_{a_{i,0}, b_{i,1}, \ldots, b_{i,n}}!$$

with

$$\begin{aligned} f(\vec{k}_i) &= \left(\vec{R}_i + M \cdot \vec{k}_i\right)^T \cdot \left(\vec{R}_i + M \cdot \vec{k}_i\right) \\ &= \vec{k}_i^T \cdot M^T \cdot M \cdot \vec{k}_i - 2 \cdot \vec{k}_i^T \cdot M^T \cdot \vec{R}_i + \vec{R}_i^T \cdot \vec{R}_i \end{aligned}$$

M^T is the transpose matrix of matrix M.

Method of least squares

Optimization calculus

The gradient of f is the vector of the derivatives of f.

$$\nabla f := \begin{pmatrix} \frac{\partial f}{\partial a_{i,0}} \\ \frac{\partial f}{\partial b_{i,1}} \\ \vdots \\ \frac{\partial f}{\partial b_{i,m}} \end{pmatrix}$$

∇f is read as 'nabla f'.

Method of least squares

Optimization calculus

Gradient of f

$$
\begin{aligned}
\nabla f &= M^T \cdot M \cdot \vec{k}_i + \vec{k}_i^T \cdot M^T \cdot M - 2 \cdot M^T \cdot \vec{R}_i \\
&= 2 \cdot M^T \cdot M \cdot \vec{k}_i - 2 \cdot M^T \cdot \vec{R}_i
\end{aligned}
$$

Method of least squares

Optimization calculus

The vector $\vec{k}_i$ that minimizes the sum of square deviations fulfils the condition:

$$
\begin{aligned}
& \nabla f && = 0 \\
\Leftrightarrow\quad & M_i^T \cdot M_i \cdot \vec{k}_i && = M_i^T \cdot \vec{R}_i \\
\Leftrightarrow\quad & \vec{k}_i && = \left(M_i^T \cdot M_i\right)^{-1} \cdot M_i^T \cdot \vec{R}_i
\end{aligned}
$$

Section 5

Quality of regression

Quality of regression

Coefficient of determination

Definition (coefficient of determination)

$$RD_i^2 = \frac{\sum_{t=1}^{m} \frac{1}{m} (\hat{R}_{i,t} - \bar{R}_i)^2}{\sum_{t=1}^{m} \frac{1}{m} (R_{i,t} - \bar{R}_i)^2}$$

- In statistics, the standard notion of the coefficient of determination is R^2 ('R-squared').
- For this lecture the notation RD_i^2 is chosen to avoid confusion with asset returns R_i.

Quality of regression

Coefficient of determination

	$\sum_{t=1}^{m} \frac{1}{m} \cdot \varepsilon_{i,t}^2 = \sum_{i=1}^{m} \frac{1}{m} \cdot (R_{i,t} - \hat{R}_{i,t})^2$	part of sample variance of $R_{i,t}$ not 'explained' by regression line
$+$	$\sum_{t=1}^{m} \frac{1}{m} \cdot (\hat{R}_{i,t} - \bar{R}_i)^2$	part of sample variance of $R_{i,t}$ 'explained' by regression line
$=$	$\mathrm{var}(R_i) = \sum_{t=1}^{m} \frac{1}{m} \cdot (R_{i,t} - \bar{R}_i)^2$	total sample variance of $R_{i,t}$

Quality of regression

Coefficient of determination

Interpretation

$$RD_i^2 = \frac{\text{sample variance of } R_{i,t} \text{ `explained' by regression line}}{\text{total sample variance of } R_{i,t}}$$

- Correspondingly: $RD_i^2 \in [0, 1]$
- The higher RD_i^2, the better the predictive nature of the linear regression model.
- $RD_i^2 = 1$ implies that the asset's idiosyncratic risk is expected to be zero.

Quality of regression

Correlation coefficient

Definition (correlation coefficient)

$$\rho_{i,F} = \frac{\mathrm{cov}(R_i, F)}{s_i \cdot s_F}$$

with

s_F sample standard deviation of F

s_i sample standard deviation of return on asset i

Quality of regression

Correlation

Interpretation

$$\text{img}(\rho_{i,F}) = [-1; 1]$$

- $\rho_{i,F} = 0$: No correlation.
- $\rho_{i,F} = 1$: Perfect positive correlation.
- $\rho_{i,F} = -1$: Perfect negative correlation.

In case of a simple linear regression:

$$RD_i = (\rho_{i,F})^2$$