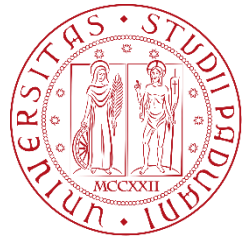


Linear Regression and the CAPM

MARINA BERTOLINI
Dipartimento di Scienze Statistiche, Università di Padova
based on the presentation on Massimiliano Caporin

The CAPM

- The Capital Asset Pricing Model (CAPM) is a financial economics model that, under a set of hypotheses, formalizes the condition of equilibrium between demand and offer of financial instruments
- The CAPM also postulates that the returns of risky investments are proportional to the returns of the *market portfolio*

$$\mu_i - r_f = \beta_i (\mu_M - r_f)$$

- The coefficient β_i measures the relation between risky assets returns (in excess of the risk-free return) and the returns of the market portfolio (in excess of the risk-free return), in *expected terms*

$$\mathbb{E}[r_{i,t}] - r_f = \beta_i (\mathbb{E}[r_{M,t}] - r_f)$$

- In this setting, to evaluate if empirical data are coherent with the model expectation, that is, the equilibrium in the market, we must first *estimate* the coefficient β_i

The CAPM

- We have to first recover the proper data, usually collected at a monthly frequency:

1. The returns of the risky assets $\rightarrow$ returns computed from equity prices
2. The risk-free return, usually the monthly return earned by investing in a government bond with short maturity
3. The market return, usually the return computed from the levels of an equity market index (a *broad market* index)

- Given these elements we can formalize the model using *time series* data

$$r_{i,t} - r_{f,t} = \alpha_i + \beta_i \left(r_{M,t} - r_{f,t}\right) + \varepsilon_{i,t}$$

where, in a linear regression model, we do have an intercept α_i, a slope coefficient β_i linking a single explanatory variable $(r_{M,t} - r_{f,t})$ to the dependent variable $(r_{i,t} - r_{f,t})$, and an error term $\varepsilon_{i,t}$

The CAPM

- The CAPM model postulates that

1. All the coefficients β_i should be different from zero and ideally positive
2. All the coefficients α_i should be equal to zero $\longrightarrow$ equilibrium condition

- In the next lectures we will discuss the statistical tools behind a linear regression model focusing on model properties, parameters estimation and inference, and some elements associated with diagnostic analysis (i.e., determine if the empirical evidence is coherent with hypotheses needed to provide proper parameters' estimation and inference)

The linear regression model

- Our objective is to evaluate the relation existing between a financial instrument and *the market* to determine which is the expected change in the financial instruments returns given a change in the market returns, namely

$$\frac{\Delta\left(r_{i,t} - r_{f,t}\right)}{\Delta\left(r_{M,t} - r_{f,t}\right)} = \frac{\Delta z_{i,t}}{\Delta z_{M,t}} = \beta_i$$

  where $z_{i,t}$ and $z_{M,t}$ are returns in excess to the risk-free returns
- How can we estimate β_i?
- Back to the sample mean...

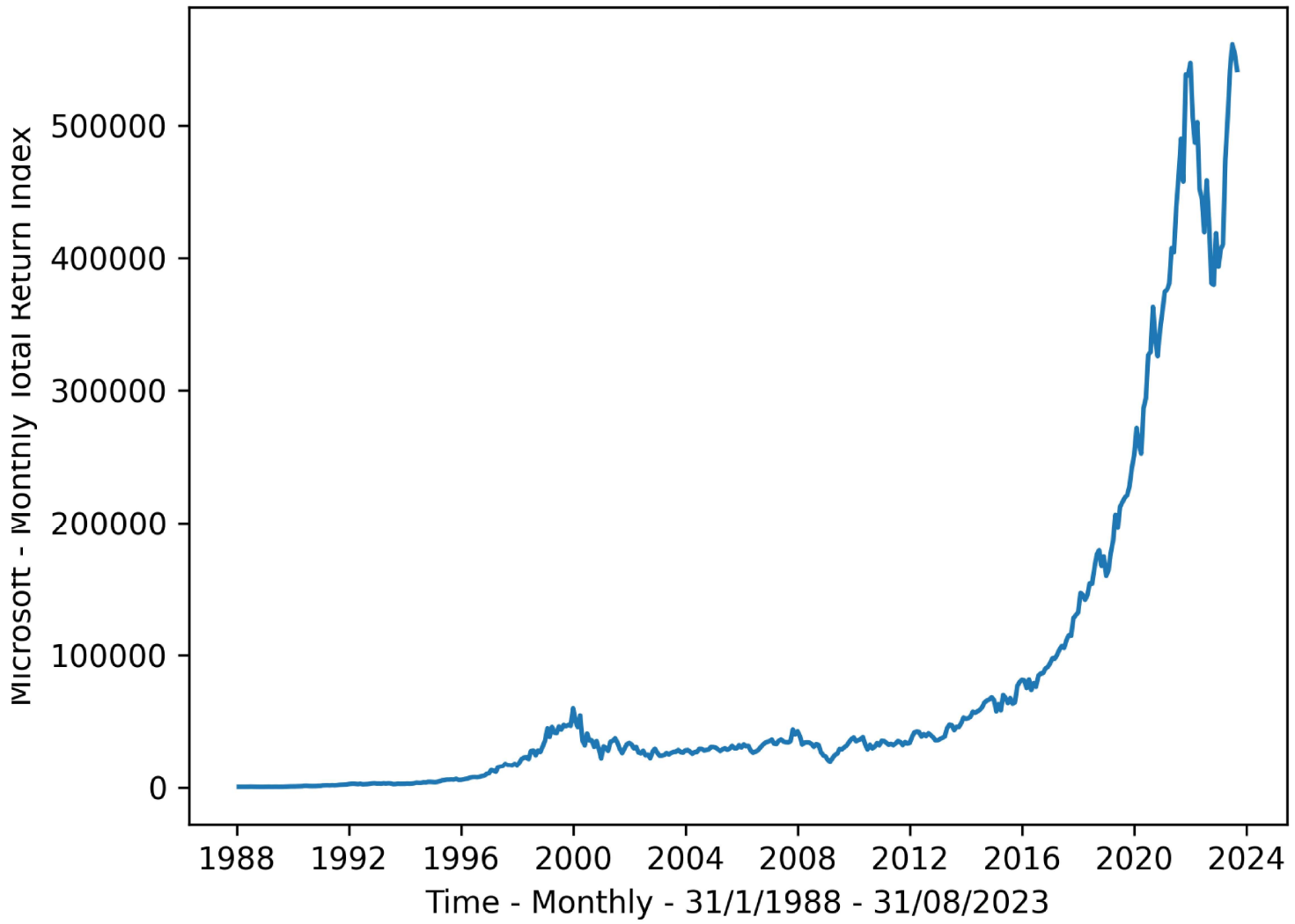
$$\bar{z}_i = \frac{1}{T}\sum_{t=1}^{T} z_{i,t}$$

...but the sample mean is the minimizer of the following problem

$$\min_{\mu} \sum_{t=1}^{T} \left(z_{i,t} - \mu\right)^2$$

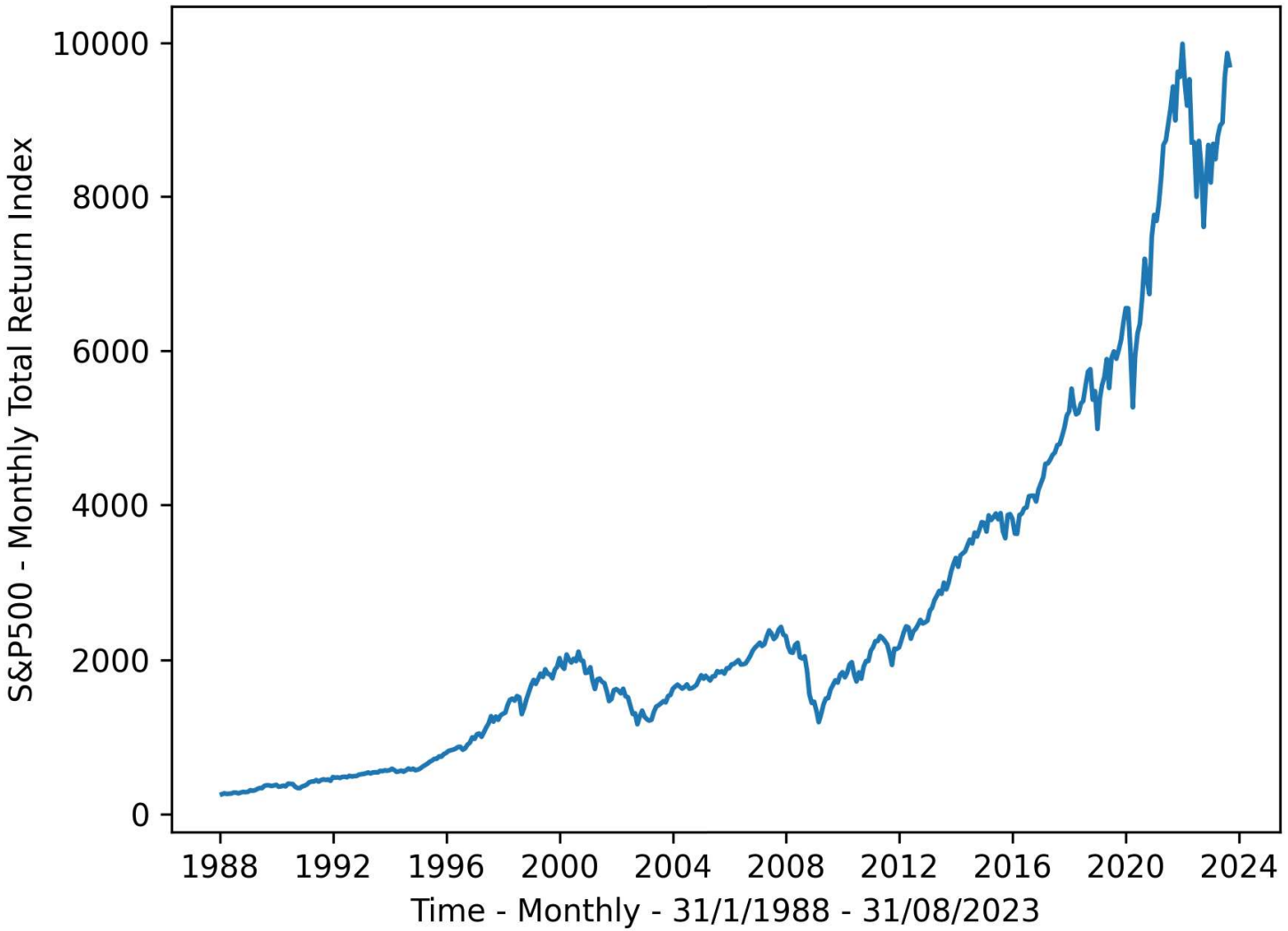
The linear regression model

Microsoft company - monthly total returns index



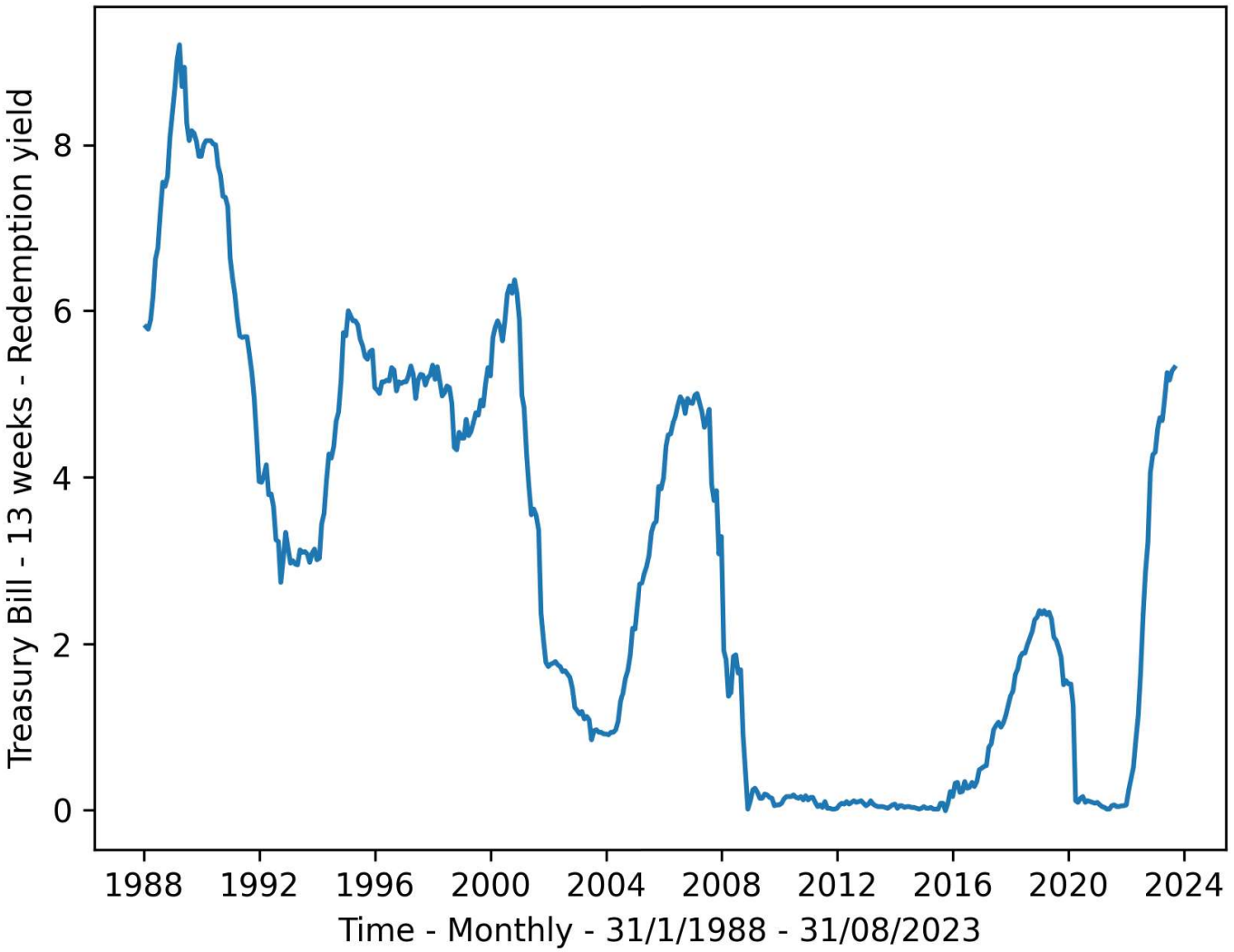
The linear regression model

S&P 500 index level (total return index)



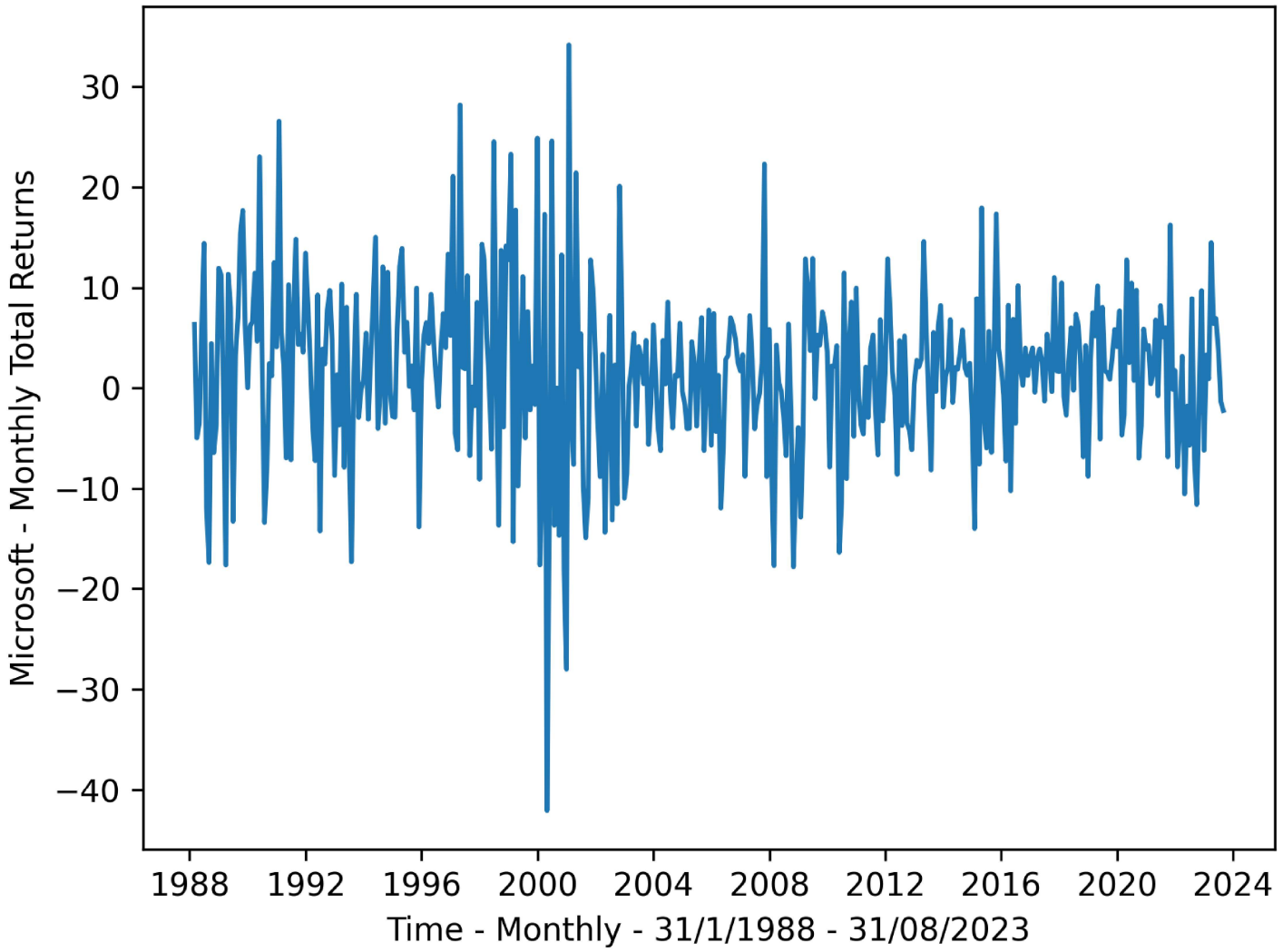
The linear regression model

Treasury Bills - 13 weeks - Annualized redemption yield - monthly frequency



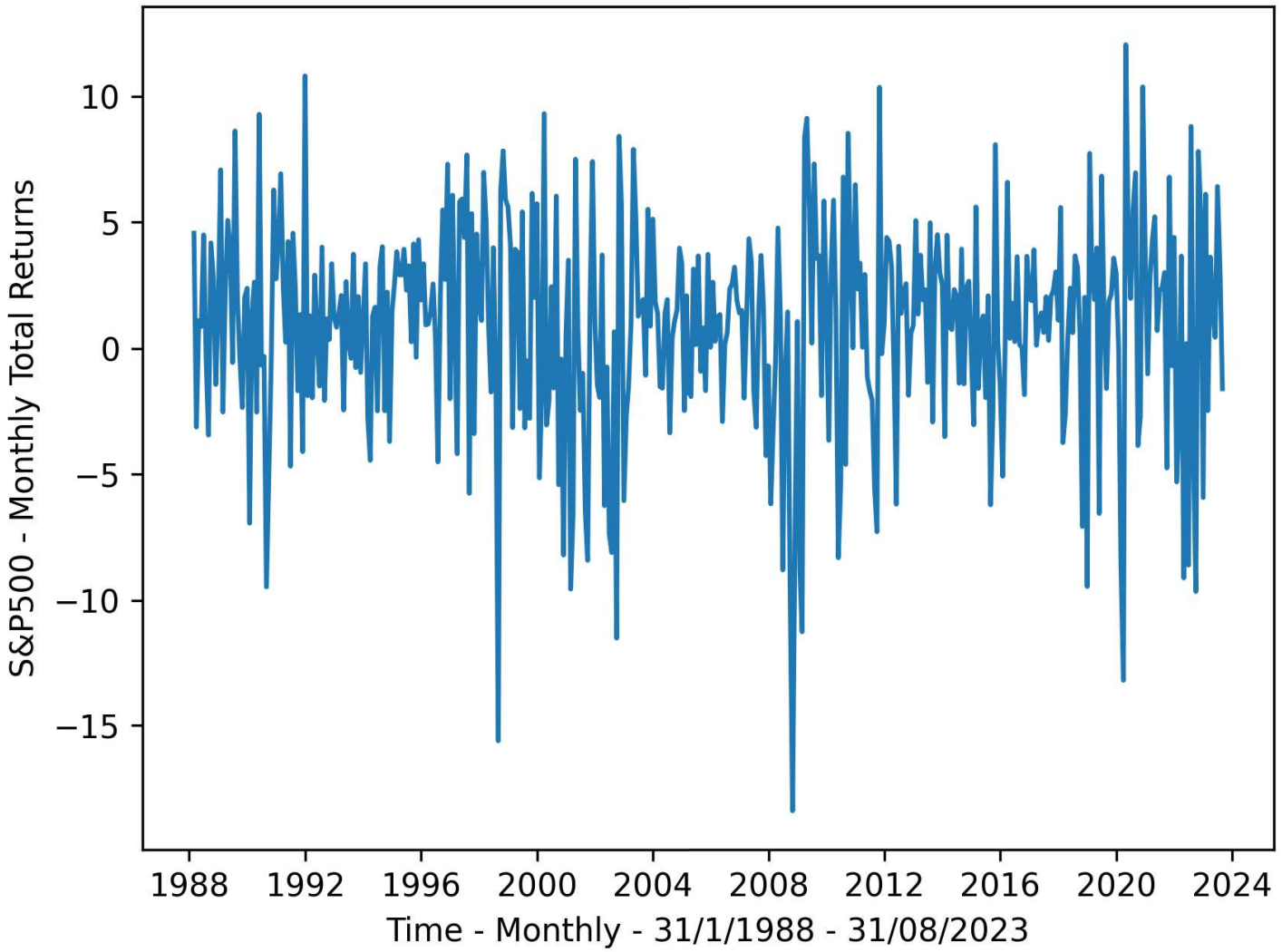
The linear regression model

Microsoft company - monthly total returns



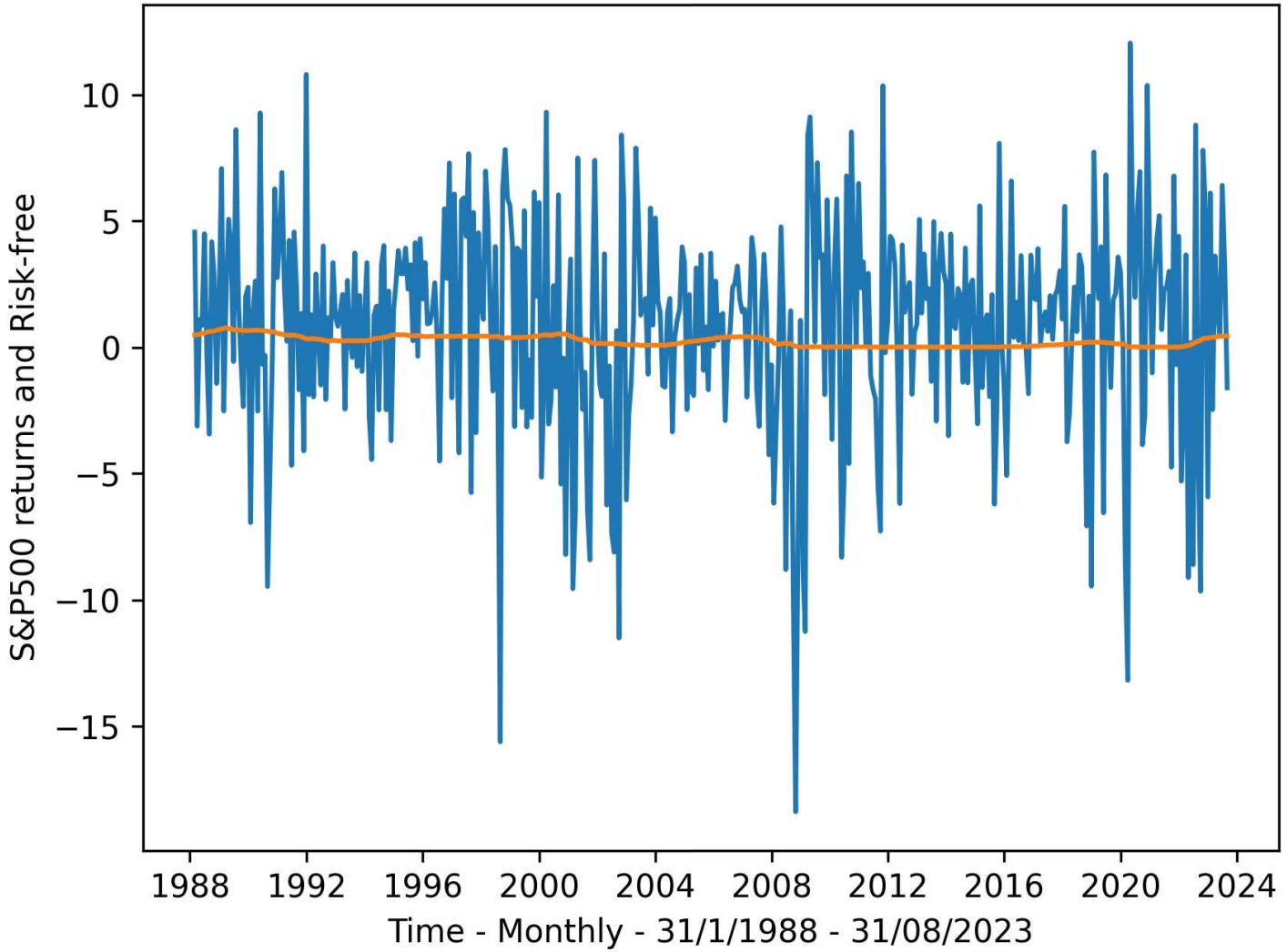
The linear regression model

S&P 500 index monthly total returns



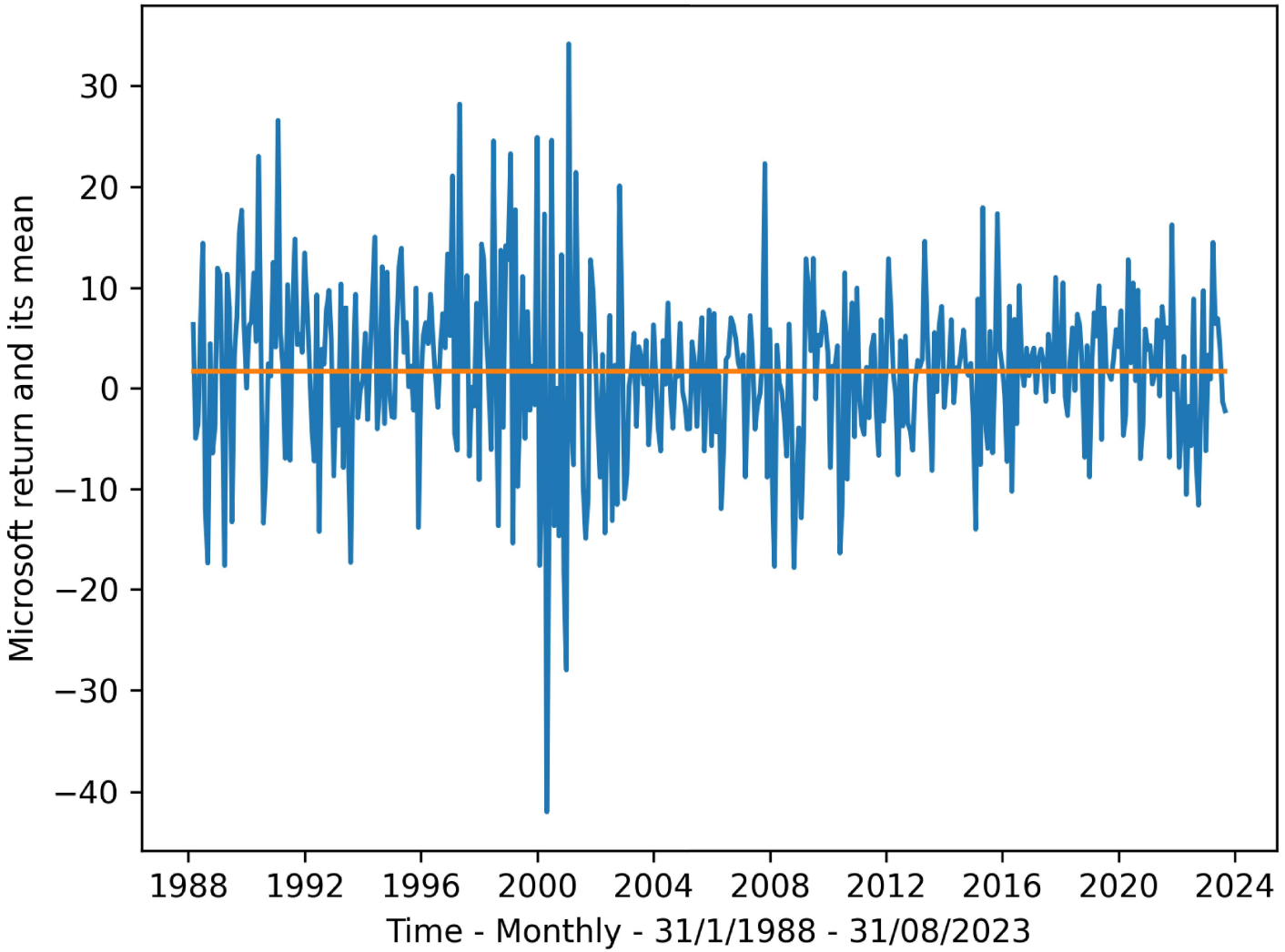
The linear regression model

S&P 500 index returns vs. Risk-free (monthly value)



The linear regression model

Microsoft total returns vs. its mean



The linear regression model

- Back to CAPM: we might represent the relation between the two variables with a scatter plot - Microsoft excess returns vs Market excess returns

The linear regression model

- The sample mean is the level that minimizes the deviation between the observations and a straight line parallel to the X axis
- With two variables, we search for the best linear approximation with the need of estimating an intercept and a slope $\rightarrow$ the problem becomes

$$\min_{\alpha_i,\beta_i} \sum_{t=1}^{T} (z_{i,t} - \alpha_i - \beta_i z_{M,t})^2$$

- We minimize the squared deviations between the observations and *predictions*, that is, the best linear approximation $\longrightarrow$ Ordinary Least Squares (OLS) estimation approach
- The analytical solutions to the minimum problem are the OLS *estimators*

The linear regression model

- The OLS estimators are

$$\hat{\beta}_i = \frac{\sum_{t=1}^{T} \left(z_{i,t} - \bar{z}_i\right)\left(z_{M,t} - \bar{z}_M\right)}{\sum_{t=1}^{T} \left(z_{M,t} - \bar{z}_M\right)^2}$$

$$\hat{\alpha}_i = \bar{z}_i - \hat{\beta}_i \bar{z}_M$$

- We also have the fitted values (best linear approximation of the relation between the stock and the market)

$$\hat{z}_{i,t} = \hat{\alpha}_i + \hat{\beta}_i z_{M,t}$$

- ...and the residuals (approximation errors)

$$\hat{e}_{i,t} = z_{i,t} - \hat{z}_{i,t} = z_{i,t} - \hat{\alpha}_i - \hat{\beta}_i z_{M,t}$$

The linear regression model

- Take the excess returns on the Microsoft stock, and regress them on the excess returns of the S&P 500 index...
- Running linear regression in Python

```
import statsmodels.api as sm
X = np.column_stack((np.ones_like(rMKTe), rMKTe))
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit()
Res1.summary()
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.332
Model:                            OLS   Adj. R-squared:                  0.330
Method:                 Least Squares   F-statistic:                     211.3
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           3.82e-39
Time:                        15:36:01   Log-Likelihood:                -1436.9
No. Observations:                 427   AIC:                             2878.
Df Residuals:                     425   BIC:                             2886.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7438      0.343      2.168      0.031       0.069       1.418
x1             1.1576      0.080     14.536      0.000       1.001       1.314
==============================================================================
```

The linear regression model

- Further output elements...

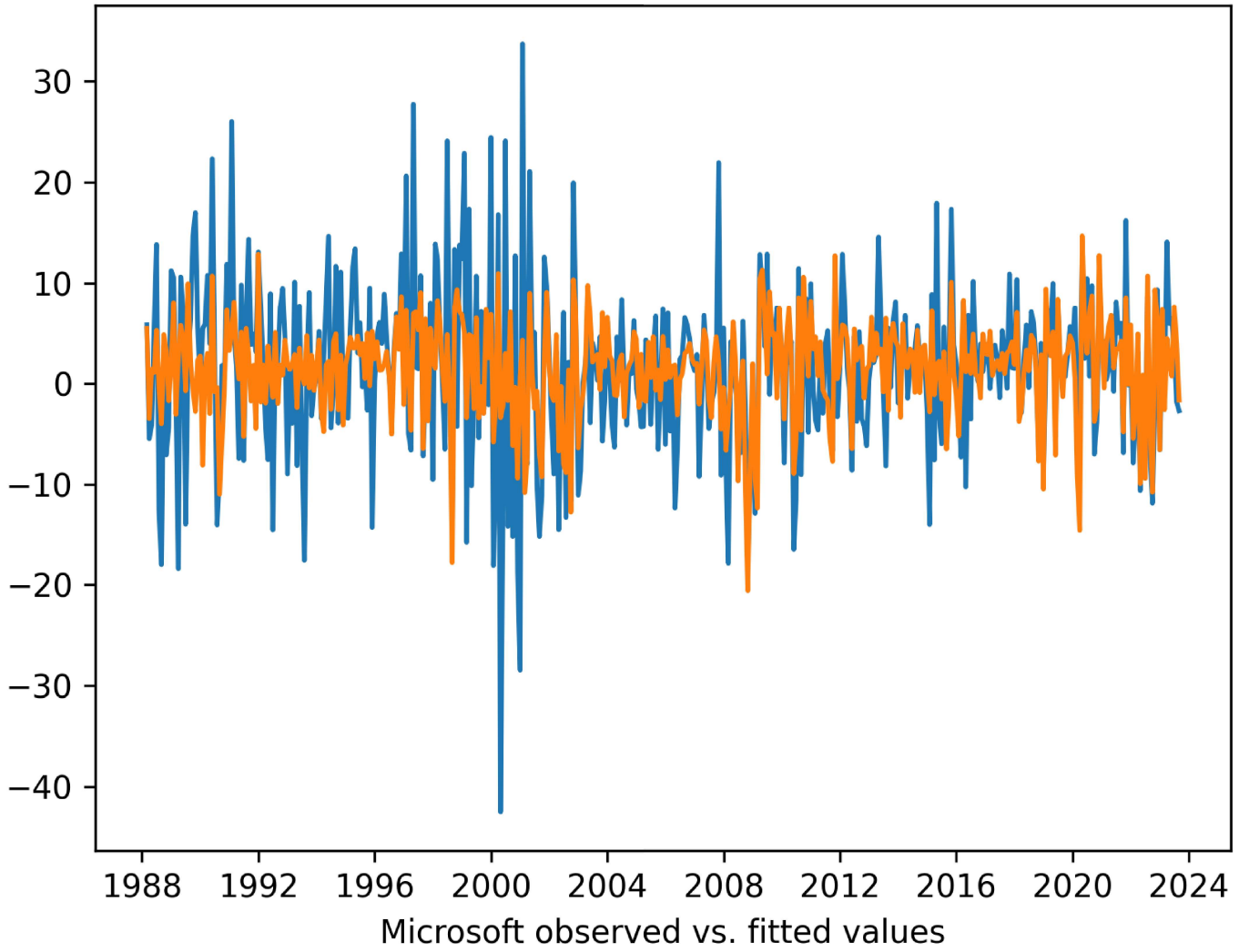
```
==============================================================================
Omnibus:                       45.093   Durbin-Watson:                   2.207
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              245.923
Skew:                          -0.182   Prob(JB):                     3.97e-54
Kurtosis:                       6.700   Cond. No.                         4.36
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the
    errors is correctly specified
```

- Single elements from the output can be selected and later used
- The output structure contains several elements to be used for additional analyses

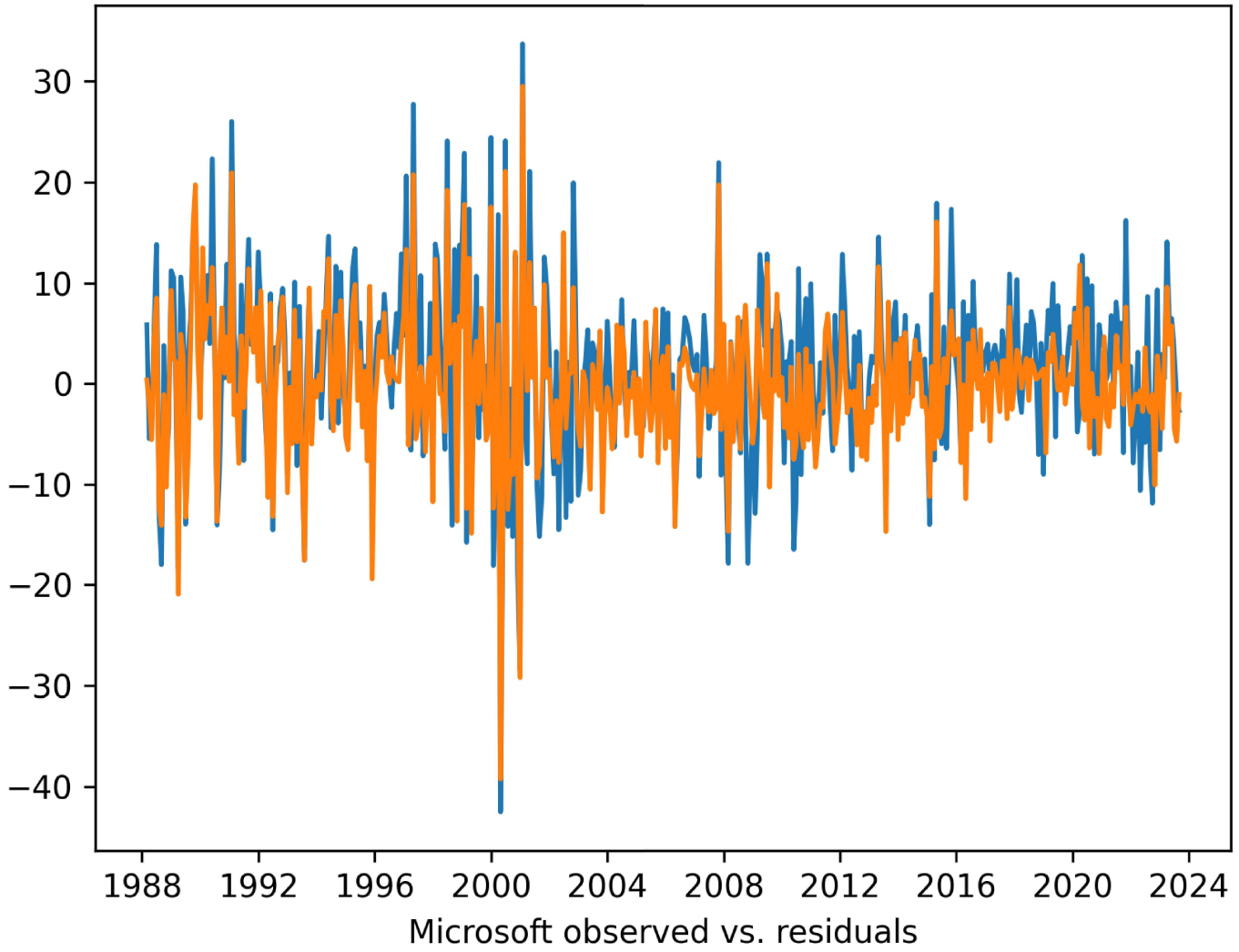
```
# recover fitted values and residuals
fit1=Res1.fittedvalues
resid1=Res1.resid
```

Microsoft total returns: observed vs. fitted values



The linear regression model

Microsoft total returns: observed vs. residuals



The linear regression model

- The estimate show that the exposure to the market index (the β_i coefficient) equals **1.158** while the intercept (the α coefficient) is **0.744**
- However, some questions emerge: a different sample period would have provided a different β, even if the model postulate a unique slope in the regression, how can we account for the uncertainty in the estimation of the parameters? how can we introduce statistical tools for determining if estimated coefficients are different from zero? can we build a confidence interval for estimated coefficients?

The linear regression model

- We must introduce some probabilistic elements
- In the previous example we used a *sample* while our interest lies in the parameters characterizing the relation between the market index and the stock returns in the *population*
- The CAPM postulates a relation in *expected* terms, that is, the expected value of stock returns *conditional* to the market return; the linear regression conditional expectation is

$$\mathbb{E}\left[z_{i,t}|z_{M,t}\right] = \alpha_i + \beta_i z_{M,t}$$

  and under equilibrium the CAPM model postulates $\alpha_i = 0$ and $\beta_i > 0$ (or better $\beta_i \neq 0$)
- We must be able to infer the values of the coefficients from the available data and to determine if they are coherent with the model expectations

The linear regression model

- To proceed, we must re-define the model

$$z_{i,t} = \alpha_i + \beta_i z_{M,t} + \varepsilon_{i,t}$$

- $z_{i,t}$ is the *dependent* variable, $z_{M,t}$ is the *explanatory* variable and $\varepsilon_{i,t}$ is the *residual* or *error* term
- α_i is the *intercept* and β_i is the *slope*
- Note that the error term might include the impact due to a number of possible issues: omitted variables and measurement errors are two notable examples
- The two model parameters α_i and β_i are unknown and must be *estimated* and given the estimates we proceed to verify *hypotheses* on their values, that is, to make *inference* on the parameters on the *population* starting from the values recovered from a *sample*

The linear regression model

- We are working under a number of hypotheses

1 The *conditional* mean of $z_{i,t}$ given $z_{M,t}$ equals $\mathbb{E}[z_{i,t}|z_{M,t}] = \alpha_i + \beta_i z_{M,t}$ and this is also identical to say $\mathbb{E}[\varepsilon_{i,t}|z_{M,t}] = 0$; we might even state, more generally, that $\mathbb{E}[z_{i,t}|Z_M = z_{M,t}] = \alpha_i + \beta_i z_{M,t}$ and $\mathbb{E}[\varepsilon_{i,t}|Z_M = z_{M,t}] = 0$ with Z_M being the set of observations for $z_{M,t}$

2 The observations (i.e., the sample) are independent $(z_{i,t}, z_{M,t})$ is independent from $(z_{i,s}, z_{M,s})$ for $s \neq t$, and they are coming from the same population; this is also stated as *i.i.d.* or *identically and independently distributed*

3 Both $z_{M,t}$ and $\varepsilon_{i,t}$ have finite fourth order moments, or $\mathbb{E}\left[z_{M,t}^4\right] < \infty$ and $\mathbb{E}\left[\varepsilon_{i,t}^4\right] < \infty$

The linear regression model

- For what concerns hypothesis 1, residuals have zero mean by construction (you can easily verify that in Python)
- Hypothesis 2 refers to the sampling scheme, that is, how observations are samples from the population; while this might be easier to understand in a cross-sectional setting, with time-series data this might not be immediate and observations are likely to be *serially correlated*
- This will create some additional complexity that will be later discussed; by now assume that, even if we use time series data, our observations are independent, or *serially uncorrelated* $(z_{i,t}, z_{M,t})$ is independent from $(z_{i,t-s}, z_{M,t-s})$ for $s \neq 0$
- Hypothesis 3, a bit more technical, is associated with the *absence* of abnormal data (outliers), and it is usually taken as given (not to be confused with the existence of distribution with fat tails - very common in finance)

The linear regression model

- Under these hypotheses we would like to provide an answer to a few questions:

1 How do we estimate α_i and β_i? $\longrightarrow$ OLS estimators

2 Which is the distribution of the estimators? that is, of $\hat{\alpha}_i$ and $\hat{\beta}_i$?

- Note that from the answer to the second question we will be able to evaluate both the expected value of the estimator, that is $\mathbb{E}\left[\hat{\beta}_i\right]$, and its variance $\mathbb{V}\left[\hat{\beta}_i\right]$
- For simplicity, we start by evaluating the moments of the estimators

The linear regression model

- Start from the expression of the OLS estimator

$$\hat{\beta}_i = \frac{\sum_{t=1}^{T} \left(z_{M,t} - \bar{z}_M\right)\left(z_{i,t} - \bar{z}_i\right)}{\sum_{t=1}^{T} \left(z_{M,t} - \bar{z}_M\right)^2}$$

- We replace $z_{i,t} - \bar{z}_i$ by first taking averages of the model $\bar{z}_i = \alpha_i + \beta_i \bar{z}_M + \bar{\varepsilon}_i$ and then subtracting the average from the model, obtaining

$$z_{i,t} - \bar{z}_i = \beta_i \left(z_{M,t} - \bar{z}_M\right) + \varepsilon_{i,t} - \bar{\varepsilon}_i$$

- We further note that $\bar{\varepsilon}_i = 0$ by construction (and coherently with Hypothesis 1) as $\hat{\alpha}_i = \bar{z}_i - \hat{\beta}_i \bar{z}_M$ and

$$\begin{aligned}
\bar{\varepsilon}_i &= \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{i,t} = \frac{1}{T}\sum_{t=1}^{T}\left(z_{i,t} - \hat{\alpha}_i - \hat{\beta}_i z_{M,t}\right) \\
&= \frac{1}{T}\sum_{t=1}^{T} z_{i,t} - \hat{\alpha}_i - \hat{\beta}_i \frac{1}{T}\sum_{t=1}^{T} z_{M,t} = \bar{z}_i - \hat{\beta}_i \bar{z}_M - \hat{\alpha}_i = 0
\end{aligned}$$

The linear regression model

- The estimator becomes

$$\begin{aligned}\hat{\beta}_i &= \frac{\sum_{t=1}^{T}\left(z_{M,t}-\bar{z}_M\right)\left(\beta_i\left(z_{M,t}-\bar{z}_M\right)+\varepsilon_{i,t}\right)}{\sum_{t=1}^{T}\left(z_{M,t}-\bar{Z}_M\right)^2} \\ &= \beta_i+\frac{\sum_{t=1}^{T}\left(z_{M,t}-\bar{z}_M\right)\varepsilon_{i,t}}{\sum_{t=1}^{T}\left(z_{M,t}-\bar{Z}_M\right)^2}\end{aligned}$$

- We are now ready to evaluate the mean and the variance
- For the mean, remind that $\mathbb{E}\left[\beta_i\right]=\beta_i$ and the law of iterated expectations allow us to focus on $\mathbb{E}\left[\hat{\beta}_i|\mathbf{Z}_M\right]$ as $\mathbb{E}\left[\hat{\beta}_i\right]=\mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_i|\mathbf{Z}_M\right]\right]$, thus

$$\mathbb{E}\left[\hat{\beta}_i|\mathbf{Z}_M\right] = \beta_i+\frac{\sum_{t=1}^{T}\left(z_{M,t}-\bar{z}_M\right)\mathbb{E}\left[\varepsilon_{i,t}|\mathbf{Z}_M\right]}{\sum_{t=1}^{T}\left(z_{M,t}-\bar{Z}_M\right)^2}$$

The linear regression model

- Given the hypotheses ($\mathbb{E}[\varepsilon_{i,t}|Z_M] = 0$) we have

$$\mathbb{E}\left[\hat{\beta}_i|Z_M\right] = \beta_i$$

and

$$\mathbb{E}\left[\hat{\beta}_i\right] = \beta_i$$

- The OLS estimator is *unbiased*
- For the intercept we have

$$\begin{aligned}\hat{\alpha}_i &= \bar{z}_i - \hat{\beta}_i \bar{z}_M \\ &= \alpha_i + \beta_i \bar{z}_M - \hat{\beta}_i \bar{z}_M \\ &= \alpha_i + \left(\beta_i - \hat{\beta}_i\right) \bar{z}_M\end{aligned}$$

- We can easily verify $\mathbb{E}[\hat{\alpha}_i] = \alpha_i$

The linear regression model

- Move now to the evaluation of the variance, starting from noticing that

$$\hat{\beta}_i - \mathbb{E}\left[\beta_i\right] = \frac{\sum_{t=1}^{T}\left(z_{M,t} - \bar{z}_M\right)\varepsilon_{i,t}}{\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2}$$

  and thus

$$\mathbb{V}\left[\hat{\beta}_i\right] = \mathbb{E}\left[\left(\hat{\beta}_i - \mathbb{E}\left[\hat{\beta}_i\right]\right)^2\right] = \mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2\right]$$

- We have to evaluate the expectation (again conditioning on Z_M and using the law of iterated expectations)

$$\mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2 | Z_M\right] = \mathbb{E}\left[\left(\frac{\sum_{t=1}^{T}\left(z_{M,t} - \bar{z}_M\right)\varepsilon_{i,t}}{\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2}\right)^2 | Z_M\right]$$

The linear regression model

- By expanding the numerator and using the linearity of the expectation operator, we have

$$\mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2 |\mathbf{Z}_M\right] = \mathbb{E}\left[\frac{\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2 \varepsilon_{i,t}^2}{\left(\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2\right)^2} |\mathbf{Z}_M\right] +$$

$$+\mathbb{E}\left[\frac{2\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left(z_{M,t} - \bar{Z}_M\right)\left(z_{M,s} - \bar{Z}_M\right)\varepsilon_{i,t}\varepsilon_{s,i}}{\left(\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2\right)^2} |\mathbf{Z}_M\right]$$

- We proceed by using again the properties of the expectation operator and the fact that $z_{M,t}$ is known given $\mathbb{Z}_M$

The linear regression model

- We obtain

$$\mathbb{E}\left[\left(\hat{\beta}_i - \beta_i\right)^2 |\mathbf{Z}_M\right] = \frac{\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2 \mathbb{E}\left[\varepsilon_{i,t}^2|\mathbf{Z}_M\right]}{\left(\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2\right)^2} +$$
$$+\frac{2\sum_{t=1}^{T-1}\sum_{s=t+1}^{T}\left(z_{M,t} - \bar{Z}_M\right)\left(z_{M,s} - \bar{Z}_M\right)\mathbb{E}\left[\varepsilon_{i,t}\varepsilon_{s,i}|\mathbf{Z}_M\right]}{\left(\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2\right)^2}$$

- To proceed we first recall the *independence* hypothesis, allowing us to state that $\mathbb{E}\left[\varepsilon_{i,t}\varepsilon_{s,i}|\mathbf{Z}_M\right] = 0$ for $t \neq s$, and then we must specify one further property of $\varepsilon_{i,t}$ related to its variance
- First recall that $\mathbb{E}\left[\varepsilon_{i,t}|\mathbf{Z}_M\right] = 0$ and thus $\mathbb{E}\left[\varepsilon_{i,t}^2|\mathbf{Z}_M\right] = \mathbb{V}\left[\varepsilon_{i,t}|\mathbf{Z}_M\right]$

The linear regression model

- For the error term, we might have two possible cases

1. Homoskedasticity: $\mathbb{V}[\varepsilon_{i,t}|Z_M] = \sigma_i^2 \longrightarrow$ all error terms have the same variance (for given i)
2. Heteroskedasticity: $\mathbb{V}[\varepsilon_{i,t}|Z_M] = \sigma_{i,t}^2 \longrightarrow$ variances change over time

- We work under case 1; for case 2 $\longrightarrow$ some insight when dealing with Time Series and than Quantitative Risk Management course
- Under homoskedasticity we have

$$\mathbb{V}\left[\hat{\beta}_i\right] = \frac{\sigma_i^2}{\sum_{t=1}^{T}\left(z_{M,t} - \bar{Z}_M\right)^2} = \frac{1}{T}\frac{\sigma_i^2}{\sigma_M^2}$$

- For the intercept

$$\mathbb{V}[\hat{\alpha}_i] = \mathbb{E}\left[(\hat{\alpha}_i - \alpha_i)^2\right] = \mathbb{E}\left[\left(\beta_i - \hat{\beta}_i\right)^2 \bar{z}_M^2\right]$$

The linear regression model

- Using previous results

$$\begin{aligned}\mathbb{V}\left[\hat{\alpha}_i\right] &= \mathbb{E}\left[\left(\beta_i - \hat{\beta}_i\right)^2 | Z_M\right] \bar{z}_M^2 \\ &= \frac{1}{T}\frac{\sigma_i^2 \bar{z}_M^2}{\sigma_M^2} = \frac{\sigma_i^2}{T}\frac{\sigma_M^2 + \bar{z}_{2,M}}{\sigma_M^2}\end{aligned}$$

- The last element we have to specify is σ_i^2 that has to be estimated
- Standard estimator is

$$\hat{\sigma}_i^2 = \frac{1}{T-2}\sum_{t=1}^{T}\hat{\varepsilon}_{t,i}^2 = \frac{1}{T-2}\sum_{t=1}^{T}\left(z_{i,t} - \hat{\alpha}_i - \hat{\beta}_i z_{M,t}\right)^2$$

corrected by $T-2$ as we estimated 2 parameters

The linear regression model

- **Gauss-Markov theorem**
  IF i) the conditional expectation is linear (as in our case as $\mathbb{E}[z_{i,t}|Z_M] = \alpha_i + \beta_i z_{M,t}$ *- equivalent to require that residuals have zero conditional mean); ii) the residuals are homoskedastic and with finite variance; iii) the residuals are serially uncorrelated (is a consequence of the independence between observations)*
  THEN
  the OLS estimator is the linear unbiased estimator with minimum variance
- The Gauss-Markov theorem, without imposing assumptions on the distribution of the data in the population provides an optimality feature of the OLS estimator, being the one with minimum dispersion for the estimators of the parameters

The linear regression model

- Large sample properties of the OLS estimator, that is for $T \to \infty$ (reported for $\hat{\beta}_i$ but holds also for $\hat{\alpha}_i$)
- **Consistency**: $\hat{\beta}_i \to_p \beta_i$, where $\to_p$ means *Convergence in probability*, even expressed as $\textit{Prob}\left(|\hat{\beta}_i - \beta_i| < \epsilon\right) \to 1$

  This is a consequence of *unbiasedness* of the estimator and of the fact that, for T diverging, the variance converges to zero
- **Asymptotic normality**: for large (diverging) T the distribution of the estimators can be approximated by (converges to) the Normal, that is

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\mathbb{V}\left[\hat{\beta}_i\right]}} \to \mathcal{N}(0,1)$$

- In small samples the distribution is a Student-T $\mathcal{T}_{T-2}$ with $T-2$ degrees of freedom

The linear regression model

- We have estimated the model parameters and we do have the informations allowing us to determine if, for asset i, the data evidence is coherent with the model, the CAPM
- Our interest is thus in the following statistical hypotheses; $H_0 : \beta_i \neq 0$ and $H_0 : \alpha_i = 0$
- In both cases, the hypothesis can be verified by means of the *test of significance* of the parameter, that is, the test verifying that the parameter is zero in the population
- Note that our expectation is for a rejection of the null hypothesis for β_i but not for α_i
- The test statistic derives from the estimator's distribution in large samples (as those typical of financial data - $T > 50$), and is

$$\frac{\hat{\beta}_i}{\sqrt{\mathbb{V}\left[\hat{\beta}_i\right]}} \rightarrow \mathcal{N}(0,1)$$

The linear regression model

- Back to our example...

```
Res1.summary()
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.332
Model:                            OLS   Adj. R-squared:                  0.330
Method:                 Least Squares   F-statistic:                     211.3
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           3.82e-39
Time:                        15:36:01   Log-Likelihood:                -1436.9
No. Observations:                 427   AIC:                             2878.
Df Residuals:                     425   BIC:                             2886.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7438      0.343      2.168      0.031      0.069      1.418
x1             1.1576      0.080     14.536      0.000      1.001      1.314
==============================================================================
```

The linear regression model

- How to summarize the fit of the model to the data? With the R^2
- The R^2 or *coefficient of determination* is a number ranging from 0 (no fit) to 1 (perfect fit) monitoring the fit of the linear model to the data
- The R^2 measures the fraction of the variance (of the dependent variable) explained by the model and corresponds to

$$R^2 = 1 - \frac{\sum_{t=1}^{T} \varepsilon_{i,t}^2}{\sum_{t=1}^{T} (z_{i,t} - \bar{z}_i)^2} = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS} = \frac{\sum_{t=1}^{T} \left(\hat{z}_{i,t} - \bar{\hat{z}}_{i,t}\right)^2}{TSS}$$

  where TSS = Total Sum of Squares, RSS = Residuals Sum of Squares and ESS = Explained Sum of Squares
- In the CAPM model (linear regression with one explanatory variable), the R^2 corresponds to the squared value of the correlation between the dependent variable (the asset returns) and the explanatory variable (the market returns)

The linear regression model

- The fit of the linear model, even in a CAPM-like setting, can sensibly change...
- When using as *dependent* the returns of a portfolio of large companies [source: Kenneth French website - using the Market factor and the returns of the 10th decile portfolio with companies sorted by market equity]

```
==============================================================
Dep. Variable:                  LARGE   R-squared:                       0.962
Model:                            OLS   Adj. R-squared:                  0.962
No. Observations:                 283   AIC:                             721.6
==============================================================
                 coef   std err          t      P>|t|    [0.025    0.975]
--------------------------------------------------------------
const          0.1047     0.052      2.027      0.044     0.003     0.206
x1             0.9379     0.011     84.264      0.000     0.916     0.960
==============================================================
```

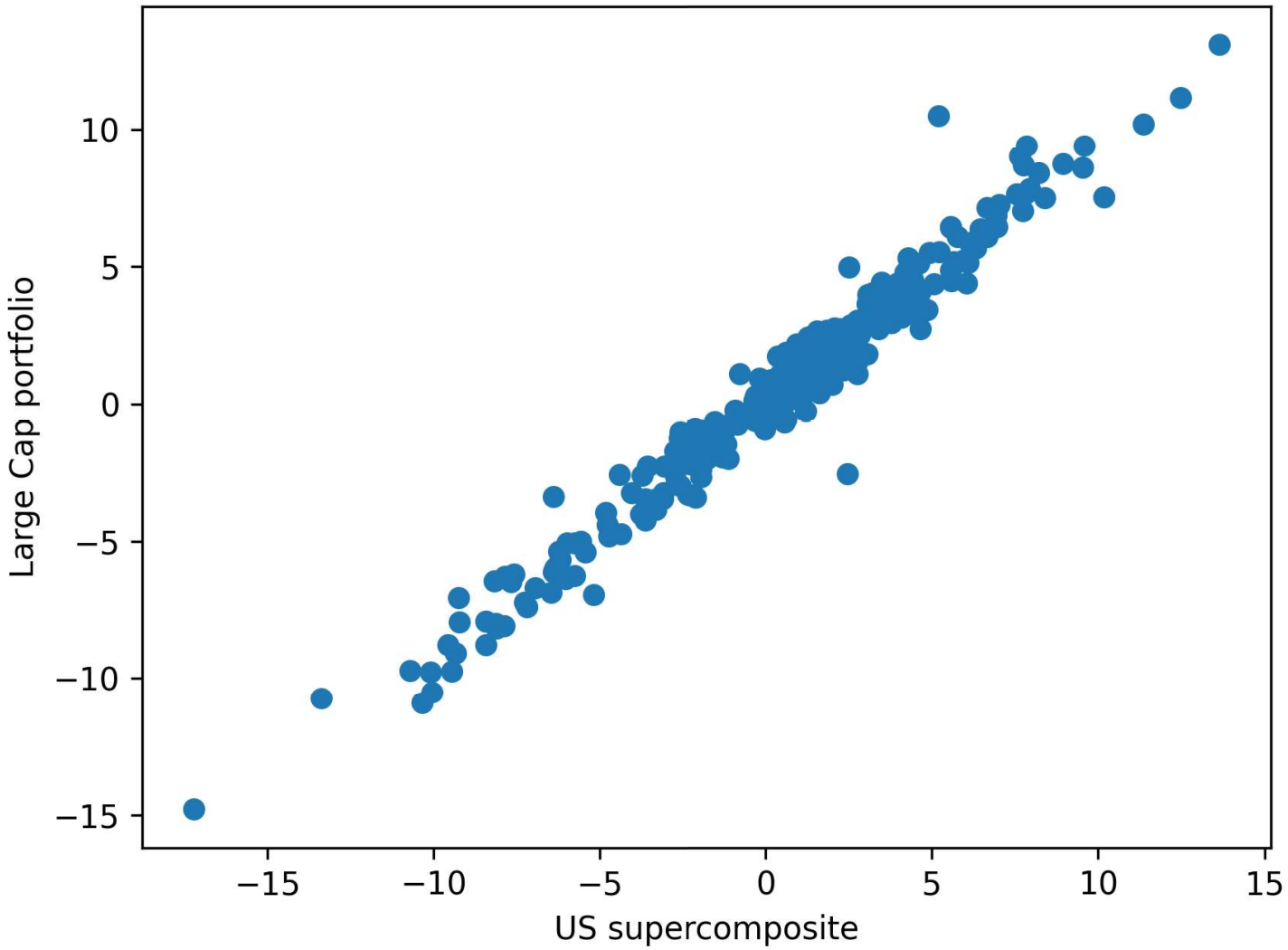
The linear regression model

- When using as *dependent* the returns of an ETF tracking the S&P500 and using as market the S&P500 index [source: Refinitiv]

```
==============================================================================
Dep. Variable: SPDR S&P 500 ETF TRUST    R-squared:                       0.994
Model:                             OLS    Adj. R-squared:                  0.994
No. Observations:                  283    AIC:                             194.3
==============================================================================
            coef   std err           t     P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      0.0001     0.020       0.007     0.994      -0.040      0.040
x1         0.9978     0.005     221.411     0.000       0.989      1.007
==============================================================================
```

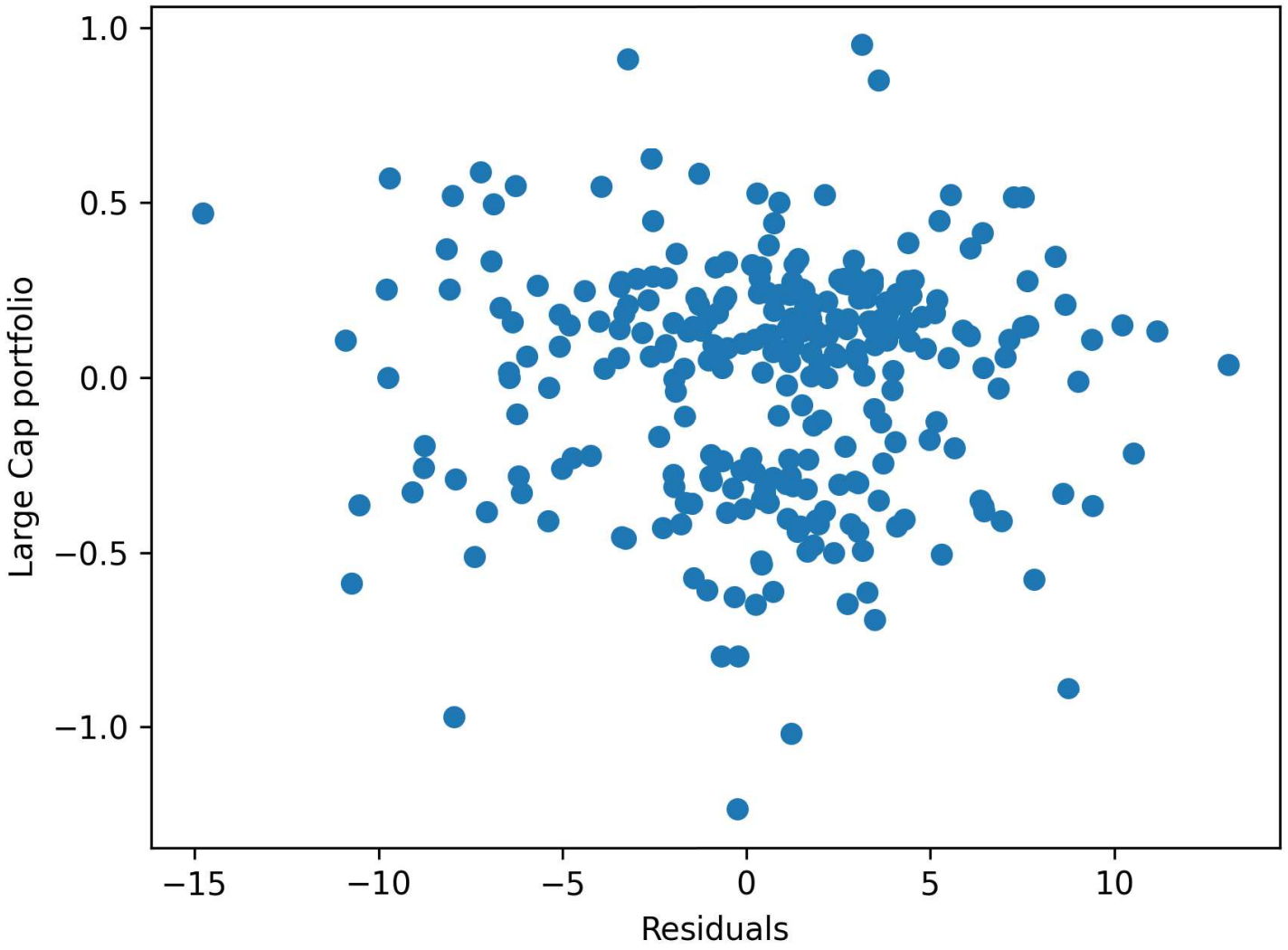
The linear regression model

Large Cap returns: market vs. observed



The linear regression model

Large Cap returns: observed vs. residuals



The linear regression model

- The linear regression model is (expectedly) more general than the simple CAPM case we considered...
- In general terms, the model might have more than a single explanatory variable, but why do we need to account for all possible (economically justified) variables that are impacting on the dependent variable? Let's consider a simple case...
- A linear regression model with two explanatory variables

$$z_{i,t} = \alpha_i + \beta_i z_{M,t} + \delta_i x_t + \nu_{i,t}$$

- Assume the previous model is the *true* one and that, making an error, we estimate the model with only one explanatory variable

$$z_{i,t} = \alpha_i + \beta_i z_{M,t} + \varepsilon_{i,t}$$

where we have $\varepsilon_{i,t} = \delta_i x_t + \nu_{i,t}$

The linear regression model

- We already know that

$$\hat{\beta}_i = \beta_i + \frac{\sum_{t=1}^{T} (z_{M,t} - \bar{z}_M) \varepsilon_{i,t}}{\sum_{t=1}^{T} (z_{M,t} - \bar{z}_M)^2}$$

  or equivalently

$$\hat{\beta}_i = \beta_i + \frac{Cov(z_{M,t}, \varepsilon_{i,t})}{\mathbb{V}[z_{M,t}]}$$

- We have seen that $\hat{\beta}_i \to_p \beta_i$ when $Cov(z_{M,t}, \varepsilon_{i,t}) \to 0$, as in the CAPM case (when the model is correctly specified)
- With two explanatory variables we have

$$Cov(z_{M,t}, \varepsilon_{i,t}) = Cov(z_{M,t}, \delta_i x_t + \nu_{i,t}) = \delta_i Cov(z_{M,t}, x_t) + Cov(z_{M,t}, \nu_{i,t})$$

- Therefore

$$\hat{\beta}_i \quad = \quad \beta_i + \delta_i \frac{Cov\left(z_{M,t}, x_t\right)}{\mathbb{V}\left[z_{M,t}\right]} + \frac{Cov\left(z_{M,t}, \nu_{i,t}\right)}{\mathbb{V}\left[z_{M,t}\right]}$$

- Under the true model (i.e., the one with two explanatory variables) the error is uncorrelated with the explanatory variables, we have

$$\hat{\beta}_i \quad \rightarrow_p \quad \beta_i + \delta_i \frac{Cov\left(z_{M,t}, x_t\right)}{\mathbb{V}\left[z_{M,t}\right]}$$

  the estimator is *inconsistent* and we have a *distortion* in the OLS estimator, that is, the estimator **does not** converge to the true value
- The distortion depends on the covariance between the *omitted* variable x_t and the explanatory variable
- We call this *distortion due to omitted variables*

The linear regression model

- The distortion size and sign depends on the covariance, with *underestimation* when $Cov\left(z_{M,t}, x_t\right) < 0$ and *overestimation* if $Cov\left(z_{M,t}, x_t\right) > 0$
- Note that the distortion *goes to* 0 even in the case $Cov\left(z_{M,t}, x_t\right) = 0$
- We have to proceed to the estimation of a model with more than one regressor...
- Linear regression model with multiple explanatory variables

$$y_t = \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} \ldots + \beta_{k-1} x_{k-1,t} + \varepsilon_t$$

- In the model the α coefficient is the *intercept* and the β_i coefficients represent the impact of the covariates (or explanatory variables) on the independent variable

The linear regression model

- Estimation is performed by OLS, that is, by minimizing

$$\min_{\alpha,\beta_1,\beta_2,\ldots,\beta_{k-1}} \sum_{t=1}^{T} (y_t - \alpha - \beta_1 x_{1,t} - \beta_2 x_{2,t} \ldots - \beta_{k-1} x_{k-1,t})^2$$

- Hypotheses adapted to multiple explanatory variables

1 The *conditional* mean of y_t given $X_1 \ldots X_{k-1}$ equals $\mathbb{E}[y_t | X_1 \ldots X_{k-1}] = \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} \ldots + \beta_{k-1} x_{k-1,t}$ and this

2 The observations (i.e., the sample) are independent

3 $x_{1,t} \ldots x_{k-1,t}$ and ε_t have finite fourth order moments

The linear regression model

- Under the three hypotheses, the OLS estimators of the parameters of interest are *consistent* and *asymptotically normal*
- Therefore, the very same statistical tools of the simpler model can be applied to verify the statistical significance of each single coefficient

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\mathbb{V}\left[\hat{\beta}_i\right]}} \rightarrow \mathcal{N}(0,1)$$

- To test $H_0 : \beta_i = 0$ we use $\hat{\beta}_i \left(\sqrt{\mathbb{V}\left[\hat{\beta}_i\right]}\right)^{-1}$
- Which is the expression of $\mathbb{V}\left[\hat{\beta}_i\right]$?
- Need to rewrite the model in matrix form...

The linear regression model

- Consider a sample of T observations and the following vectors and matrices

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & \dots & x_{1,k-1} \\ 1 & x_{2,1} & \dots & x_{2,k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T,1} & \dots & x_{T,k-1} \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{bmatrix}$$

- The model becomes

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

- Hypothesis 2) might be recast in $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}_T$ which implies that error terms are independent (serially uncorrelated) and **also** homoskedastic as $\mathbb{V}[\varepsilon_t] = \sigma^2$ for all t and $\mathbb{E}[\varepsilon_t \varepsilon_s] = 0$ for all $t \neq s$

The linear regression model

- OLS estimation minimizes

$$\min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

- Imposing first order conditions for minimization leads to

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}'} = 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{X}'\boldsymbol{y} = 0$$

- Then, the OLS estimator **of the vector** $\boldsymbol{\beta}$ is

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

  which exist if $\boldsymbol{X}$ has *full column rank* $\rightarrow$ absence of collinearity in explanatory variables

The linear regression model

- Under the homoskedasticity assumption, and given the independence assumption between explanatory variables and the error term, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right) \\
&= \boldsymbol{\beta}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\
\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \beta \\
\mathbb{V}\left[\hat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{2}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\
&= \sigma^{2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}
\end{aligned}$$

- Therefore, we have proved *unbiasedness* and we also have consistency as, for increasing T, $\boldsymbol{X}'\boldsymbol{X}$ explodes

The linear regression model

- Back to our question... the variance of one single coefficient...

$$\mathbb{V}\left[\hat{\beta}_i\right] = \sigma^2 a_{ii}$$

where a_{ii} is the diagonal element of position i in $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, and the variance of the error term is estimated as

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{T-k}$$

- In addition, the entire vector $\boldsymbol{\beta}$ follows an asymptotically normal distribution

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

The linear regression model

- The knowledge of the distribution for the entire vector of parameters allows us to dig further in testing hypotheses on coefficients, namely hypotheses involving more than a single coefficient
- Two possible ways: i) testing *linear* restrictions on coefficients; ii) estimating a restricted model (where restrictions have been applied to the equations before estimation)
- Consider the first, more general, case; the linear restrictions hypothesis can be casted in the following system

$$H_0 : \mathcal{R}\beta = r$$

where $\mathcal{R}$ is the $q \times k$ *selection* matrix (the matrix implementing the linear relations across parameters) and r is the $q \times 1$ *target* vector (the value of parameters under the restrictions)

- As an example consider $k = 5$ and the restriction $\beta_1 + \beta_2 = \beta_4$; the system becomes

$$[0 \quad 1 \quad 1 \quad 0 \quad -1] \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} = 0$$

- Note that q is the number of restrictions
- A special case is the test of *significance of the entire regression*, that is, we test all betas to be jointly equal to zero (we exclude the intercept)

$$\mathcal{R} = [0_{k-1} | \boldsymbol{I}_{k-1}], \quad r = 0_{q-1}$$

The linear regression model

- How to verify the null hypothesis? Back to properties of the Normal...
- For a $k-$dimensional Normal random variable $\boldsymbol{x}$ it holds that

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \Rightarrow (\boldsymbol{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \sim \chi^2_k$$

- Therefore, in our case, we have

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

$$\mathcal{R}\hat{\boldsymbol{\beta}} - r \sim \mathcal{N}\left(\mathcal{R}\boldsymbol{\beta} - r, \sigma^2 \mathcal{R} (\boldsymbol{X}'\boldsymbol{X})^{-1} \mathcal{R}'\right)$$

$$\left(\mathcal{R}\hat{\boldsymbol{\beta}} - r\right)' \left[\sigma^2 \mathcal{R} (\boldsymbol{X}'\boldsymbol{X})^{-1} \mathcal{R}'\right]^{-1} \left(\mathcal{R}\hat{\boldsymbol{\beta}} - r\right) \sim \chi^2_q$$

- The residuals variance is estimated in the usual way
- In small samples the statistic follows an $\mathcal{F}_{q,T-k}$ distribution and is usually referred to as *F-test*

The linear regression model

- The test of significance of the entire regression is usually included in the default outputs of implementation of linear regression (together with statistics for significance of single coefficients and the R^2)
- The test for linear restrictions can be performed according to an alternative procedure, contrasting two models
- Back to our example with $k = 5$, where we have the general, or *unrestricted* model

$$y_t = \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \beta_4 x_{4,t} + \varepsilon_t$$

- Using the restriction $\beta_1 + \beta_2 = \beta_4$ we might replace β_4 and rewrite the model as

$$\begin{aligned} y_t &= \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + \beta_4 x_{4,t} + \varepsilon_t \\ &= \alpha + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \beta_3 x_{3,t} + (\beta_1 + \beta_2) x_{4,t} + \varepsilon_t \\ &= \alpha + \beta_1 (x_{1,t} + x_{4,t}) + \beta_2 (x_{2,t} + x_{4,t}) + \beta_3 x_{3,t} + \varepsilon_t \end{aligned}$$

The linear regression model

- Therefore, we obtain the *restricted* model

$$y_t = \alpha + \beta_1 (x_{1,t} + x_{4,t}) + \beta_2 (x_{2,t} + x_{4,t}) + \beta_3 x_{3,t} + \varepsilon_t$$

- To build the test statistic we separately estimate the *unrestricted* and the *restricted* models and store their residuals, form which we compute the residuals sums of squares for the *unrestricted* model, RSS_U, and for the *restricted* model, RSS_R
- The test statistic equals

$$\frac{(RSS_R - RSS_U)/q}{RSS_U/(T-k)} \sim F_{q,T-k}$$

- In large samples $qF_{q,T-k} \sim \chi^2_q$

The linear regression model

- Note that the χ^2 (as well as the $\mathcal{F}$) distribution is defined on the positive real line; this is coherent with the construction of the F-test we have just seen as it holds that $RSS_R > RSS_U$
- This also implies that $R_R^2 > R_U^2$ irrespective of the validity of the restrictions imposed
- This holds, in particular, if restrictions point at excluding a subset of the explanatory variables, implying that, if we introduce irrelevant *regressors* the R^2 will anyway increase...
- Fortunately, these exist an *Adjusted* R^2

$$\bar{R}^2 = 1 - \frac{T-1}{T-k-1}\frac{RSS}{TSS} < R^2$$

- The Adjusted R^2 does not necessarily increase after the inclusion of an additional variable in the linear regression model

The linear regression model

- Consider a more general linear regression model inspired to the CAPM: in this case we use data of several risk factors, all recovered from the website of Kenneth French
- Data are monthly from July 1963 to July 2023
- Variables (all returns) are: the large cap (10th decile by market equity) portfolio; the Small-Minus-Big (SMB) size factor; the High-Minus-Low (HML) Book-to-Market factor; the Momentum (MOM) factor; the Robust-Minus-Weak (RMW) operating profitability factor; the Conservative-Minus-Aggressive (CMA) investment portfolios factor; the Market factor
- Data are available in the Moodle
- In the next slide the estimation output

The linear regression model

Dep. Variable:	LARGE	R-squared:	0.985
Model:	OLS	Adj. R-squared:	0.985
Method:	Least Squares	F-statistic:	7948.
Date:	Wed, 13 Sep 2023	Prob (F-statistic):	0.00
Time:	10:32:39	Log-Likelihood:	-549.49
No. Observations:	721	AIC:	1113.
Df Residuals:	714	BIC:	1145.
Df Model:	6		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3825	0.021	18.612	0.000	0.342	0.423
Market	0.9780	0.005	198.971	0.000	0.968	0.988
SMB	-0.2646	0.007	-37.270	0.000	-0.279	-0.251
HML	-0.0060	0.009	-0.637	0.524	-0.024	0.012
RMW	0.0362	0.010	3.765	0.000	0.017	0.055
CMA	-0.0080	0.014	-0.580	0.562	-0.035	0.019
MOM	-0.0002	0.005	-0.034	0.973	-0.010	0.009

Omnibus:	26.904	Durbin-Watson:	1.405
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.427
Skew:	0.449	Prob(JB):	4.07e-07
Kurtosis:	3.415	Cond. No.	5.33

The linear regression model

- The model fit to the data is impressive, $R^2 = 0.985$, but not all parameters are statistically significant
- This might lead to an impact on inferential procedures: the inclusion of irrelevant variables does not lead to biases on the relevant variables estimated parameters but generally lead to an increase in the standard errors
- Assuming that the model satisfies the needed assumptions, we can use tests for significance of parameters to perform *variables selection*
- We can proceed iteratively, removing one variable at a time, starting from those with less significant coefficients (i.e., largest p-value); this is particularly useful when the degrees of freedom are relatively small (not our case)
- Alternatively, we might perform a joint test for linear restrictions on parameters, setting all the non-significant ones to zero
- Start from the first case...

The linear regression model

Dep. Variable:	LARGE	R-squared:	0.985
Model:	OLS	Adj. R-squared:	0.985
Method:	Least Squares	F-statistic:	9551.
Date:	Wed, 13 Sep 2023	Prob (F-statistic):	0.00
Time:	10:54:13	Log-Likelihood:	-549.50
No. Observations:	721	AIC:	1111.
Df Residuals:	715	BIC:	1138.
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3824	0.020	18.857	0.000	0.343	0.422
MKT	0.9780	0.005	201.636	0.000	0.969	0.988
SMB	-0.2647	0.007	-37.309	0.000	-0.279	-0.251
HML	-0.0059	0.009	-0.651	0.515	-0.024	0.012
RMW	0.0361	0.010	3.781	0.000	0.017	0.055
CMA	-0.0081	0.014	-0.590	0.555	-0.035	0.019

Omnibus:	26.885	Durbin-Watson:	1.406
Prob(Omnibus):	0.000	Jarque-Bera (JB):	29.400
Skew:	0.449	Prob(JB):	4.13e-07
Kurtosis:	3.414	Cond. No.	5.18

The linear regression model

```
==============================================================================
Dep. Variable:                  LARGE   R-squared:                       0.985
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                 1.195e+04
Date:                Wed, 13 Sep 2023   Prob (F-statistic):               0.00
Time:                        11:42:19   Log-Likelihood:                -549.67
No. Observations:                 721   AIC:                             1109.
Df Residuals:                     716   BIC:                             1132.
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3803      0.020     19.057      0.000       0.341       0.419
MKT            0.9789      0.005    212.074      0.000       0.970       0.988
SMB           -0.2643      0.007    -37.427      0.000      -0.278      -0.250
HML           -0.0096      0.007     -1.439      0.150      -0.023       0.003
RMW            0.0371      0.009      3.953      0.000       0.019       0.056
==============================================================================
Omnibus:                       27.508   Durbin-Watson:                   1.406
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               30.185
Skew:                           0.454   Prob(JB):                     2.79e-07
Kurtosis:                       3.424   Cond. No.                         4.91
==============================================================================
```

The linear regression model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  LARGE   R-squared:                       0.985
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                 1.591e+04
Date:                Wed, 13 Sep 2023   Prob (F-statistic):               0.00
Time:                        11:43:08   Log-Likelihood:                -550.71
No. Observations:                 721   AIC:                             1109.
Df Residuals:                     717   BIC:                             1128.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.3772      0.020     18.998      0.000       0.338       0.416
MKT            0.9803      0.005    216.914      0.000       0.971       0.989
SMB           -0.2650      0.007    -37.578      0.000      -0.279      -0.251
RMW            0.0362      0.009      3.856      0.000       0.018       0.055
==============================================================================
Omnibus:                       27.641   Durbin-Watson:                   1.411
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               30.244
Skew:                           0.458   Prob(JB):                     2.71e-07
Kurtosis:                       3.408   Cond. No.                         4.83
==============================================================================
```

The linear regression model

- Removing the irrelevant variables has marginal effects on i) the model fit to the data, ii) the estimated coefficients, and in this specific case, also on the parameters' standard errors
- This result is not general, there are cases in which, by iteratively excluding irrelevant variables, standard error change in a sizeable way, impacting on the statistical significance of variables (i.e., non-significant ones might become significant, marginally significant ones might become non-significant)
- As an alternative, we can build the F-stat for joint test of zero-restrictions by comparing the R-squared of the unrestricted and restricted models (killing in one shot all non-significant parameters); in this case the P-value is 0.491 thus we do not reject the null of zero restrictions and we select the restricted model

The linear regression model

```
# load data of one worksheet
data = pd.read_excel('ExampleMultipleRegression.xlsx', 'data',
     index_col=None, na_values=['NA'])
F1=data.MKTminRF
F2=data.SMB
F3=data.HML
F4=data.RMW
F5=data.CMA
F6=data.MOM
Y=data.LARGE
X = np.column_stack((np.ones_like(F1), F1, F2, F3, F4, F5, F6))
Res1 = sm.OLS(Y, X).fit()
Res1.summary()
X = np.column_stack((np.ones_like(F1), F1, F2, F4))
Res2 = sm.OLS(Y, X).fit()
Res2.summary()
RSSU=Res1.ssr
RSSR=Res2.ssr
Fstat=((RSSR-RSSU)/3)/(RSSU/(714))
Pval=1-sp.stats.f.cdf(Fstat,3,714)
```

The linear regression model

- Using Adjusted R^2 is common, but other indicators monitoring the fit to the data are available
- A class of them is given by the *Information Criteria* that combine the *RSS* measure of the fit with the number of parameters (or variables) included in the model
- Information Criteria (IC) have the following general form

$$IC = f(RSS) + g(k)$$

- The two most common IC are

$$\begin{aligned} AIC &= \log\left(\frac{RSS}{T}\right) + 2\frac{K}{T} \\ BIC &= \log\left(\frac{RSS}{T}\right) + \log(T)\frac{K}{T} \end{aligned}$$

The linear regression model

- Information criteria increase with a worsening of the fit and increase with an increase in parameters
- Consequently, the better models (specifications) are those with *smaller* value for the IC
- AIC (Akaike Criterion) is less conservative than BIC (Bayesian or Schwartz Criterion) as $\log(T) > 2$ for usual time series length in finance

The linear regression model

- A side note: variables selection can be also done with other approaches, which are becoming common in particular when the number of potentially relevant variables is **large**
- I refer to automatic selection (generally based on information criteria), machine learning-based approaches (penalization, regularization, network-based or AI-based), or dimension reduction methods
- I'm not stating these should not be used, but I stress that a relevant aspect is the economic intuition behind the selected variables: once you introduce a variable in the model you should be able to attach an economic reason for the variable being present and to provide an economic interpretation to the coefficient (to its sign and size)
- If you fail in this respect, your approach will be a *black-box* just providing a fit (or a prediction) of the variable of interest (and this might be enough in some frameworks and depending on the objective of the modelling step)

Diagnostic analysis

- We have seen how to estimate a linear regression model and how to perform inference on the parameters
- But how can we be sure that the model we estimated has been correctly specified?
- There are several possible issues related to specification errors
- We have already encountered one case, that of *omitted variables*, which might lead to distortions in the estimated parameters (they converge to wrong values)
- Properly specifying the model is crucial to run appropriate inference and then to safely interpret the model outcome (in terms of data fit and economic intuition we might recover from the parameters)

Diagnostic analysis

- Specification errors might impact on a number of elements related to the model:
  - ✓ The explanatory variables: collinearity, omitted variables, irrelevant variables
  - ✓ The functional form: is linearity appropriate?
  - ✓ The stability of the functional form: are parameters subject to changes? (structural break)
  - ✓ The assumptions placed on the innovations (homoskedasticity and independence)
- We will address those elements and discuss how we could identify specification issues with proper statistical tests, called *diagnostic tests*

Diagnostic analysis

- **Multicollinearity** refers to the possible presence of **exact** linear relationships among the explanatory variables
- In this case, estimation by OLS cannot be performed as $\boldsymbol{X}'\boldsymbol{X}$ is not invertible; we are in the case of *perfect collinearity*
- This issue is solved by removing one of the linearly dependent variables (some software do that by default); which variable to remove? Irrelevant, they are linearly related, we do not loose information
- Typical case: the **dummy trap**
- Dummy variables are variables taking two possible values 0 or 1 according to a feature they track: gender (male= 0, female= 1), weekend (0 if weekday, 1 if weekend), seasonal dummies (for quarters, months, day of the week)

Diagnostic analysis

- We might use monthly dummies to determine, for instance, if the alpha is difference among months
- Set $d_{t,j}$ to be the dummy variables for month j, with 1 being January and 12 being December

$$d_{t,j} = \begin{cases} 0 & month(t) \neq j \\ 1 & month(t) = j \end{cases}$$

- The model will include all dummies but one...

$$z_t = \alpha + \sum_{j=1}^{11} \delta_j d_{t,j} + \beta z_{M,t} + \varepsilon_t$$

- This is need due to the presence of the constant: if we do not exclude one dummy we have $1 = \sum_{j=1}^{12} d_{t,j}$, that is, perfect collinearity

Diagnostic analysis

- In the model we consider, α corresponds to the intercept for the month we exclude from the dummy set, that is, the one where all dummies will have a zero value
- Differently, the intercept for month *j* will be equal to $\alpha + \delta_j$, and testing for a significant change in the alpha for a given month will be associated with a test on each single δ_j coefficient
- Alternatively to the exclusion of one dummy one might exclude the intercept from the model, thus estimating month-specific alphas
- Perfect (exact) collinearity is easily handled, more difficult the case of *near collinearity* that emerges when correlation among explanatory variables is high

Diagnostic analysis

- Near collinearity has consequences on both parameters estimation and inference
- Under near collinearity...
- ...we might have some parameters close one to the other but with opposite signs...
- ...standard errors being very high...
- ...large impact on inferential procedures
- How to solve this? Control for the correlation across variables and exclude variables highly correlated with other variables included in the set of explanatory

Diagnostic analysis

- To shed further light on multicollinearity, consider another representation of the variance of the estimated linear regression paramters, $\hat{\beta}_i$; it can be shown that

$$\mathbb{V}\left[\hat{\beta}_i\right] = \sigma^2 a_{ii} = \frac{\sigma^2}{\sum_{t=1}^{T}\left(x_{i,t} - \bar{x}_i\right)^2} \frac{1}{1 - R_{i.}^2}$$

  where $\sum_{t=1}^{T}\left(x_{i,t} - \bar{x}_i\right)^2 = T\mathbb{V}\left[x_{i,t}\right]$, and $R_{i.}^2$ is the R-squared in the regression of $x_{i,t}$ on all the other explanatory variables

- It follows that
  - ✓ The larges the variance of the explanatory variable, the smaller the variance of the coefficient
  - ✓ The smaller the variance of the error (i.e., the better the fit of the model), the smaller the variance of the coefficient
  - ✓ The larger the correlation of $x_{i,t}$ with other variables, the larger the variance of the coefficient

Diagnostic analysis

- We thus have a clear impact of collinearity (i.e., large correlation across explanatory variables) and the variance of the estimated coefficients
- An index used for detecting collinearity, that might be used in addition to the sample correlation across explanatory variables, is the so-called Variance Inflation Factor (VIF), that is, the second term in the coefficient variance

$$VIF_i = \frac{1}{1 - R_{i.}^2}$$

- The larger the VIF for a variable, the larger the possible impact due to collinearity; different studies propose different threshold values to be used for detecting potential issues (common choice 10)
- Other solutions beside dropping variables are: ridge regression and the use of Principal Components of the explanatory variables

Diagnostic analysis

- As anticipated, a second diagnostic analysis focuses on the functional form of the model challenging the linearity hypothesis
- In this case, our model has been estimated as

$$z_t = \alpha + \beta z_{M,t} + \varepsilon_t$$

  and we are interested in verifying the appropriateness of the linear specification
- To that purpose we run a test, called **Ramsey** test or **RESET** (Regression Equation Specification Error Test), which is using an *auxiliary* regression (i.e., a regression needed only for testing purposes and not intended to provide an economic rationale)
- The system of hypotheses is

$$\begin{cases} H_0 : \mathbb{E}\left[z_t | z_{M,t}\right] = \alpha + \beta z_{M,t} \\ H_1 : \mathbb{E}\left[z_t | z_{M,t}\right] \neq \alpha + \beta z_{M,t} \end{cases}$$

Diagnostic analysis

- To verify the null, we run the regression

$$z_t = \alpha + \beta z_{M,t} + \sum_{j=2}^{p} \delta_j \hat{z}_t^j + \nu_t$$

  where $\hat{z}_t$ is the fitted value under the linear specification
- We verify the null hypothesis by testing that the δ_j coefficients are all jointly equal to zero
- The test corresponds to a test for restrictions on coefficients, and equals

$$\frac{(RSS_L - RSS_{NL})/(p-1)}{RSS_{NL}/(T-k-p-1)} \sim F_{p-1,T-k}$$

  with RSS_L being measured on the linear model and RSS_{NL} obtained from the auxiliary regression
- For the RESET test, we usually set p to be, at maximum, equal to 3 or 4

Diagnostic analysis

- If we reject the null, we do have evidence supporting the existence of non-linearities in the relation between the variables (dependent and explanatory)
- To avoid this problem we might use a more flexible functional form: introduce powers of the explanatory, transform the variables (exponentials, logs, square root Box-Cox...)
- Note that non-linearities might also be the consequence of omitted variables and/or of structural breaks in the relation across variables

Diagnostic analysis

- As a first example consider a general production function

$$y = \alpha x_1^{\beta} x_2^{\delta}$$

  with Y being the product and x_1, x_2 being the production factors
- In this model, we might regress y on x_1 and x_2 but we will likely find evidence of non-linearities
- In fact, it would be more appropriate to model logarithms, as we have

$$\log(y) = \alpha + \beta \log(x_1) + \delta \log(x_2)$$

Diagnostic analysis

- A second example consider financial data, in details portfolio excess returns

$$z_{p,t} = \alpha + \beta z_{M,t} + \delta z_{M,t}^2 + \nu_t$$

  where the introduction of squared market excess returns would capture non-linear impacts of the latter on the returns of a portfolio
- This analysis has a clear economic interpretation, and refers to the so-called *Market Timing* models; market timing is the ability of a manager to anticipate market movements, and relates to the tendency of reducing market exposure before market contractions and augmenting market exposure before rallies

Diagnostic analysis

- Diagnostic checks on linear regression models should also focus on the stability of the relation among variables, that is, on the stability of the coefficients
- in fact, there are events that could impact on the relation among variables, with the consequence of changes in the slope of the regression, that is, on the coefficients
- Events might be market-wide, such as a crisis, the adoption of new regulations, or company-specific, as a new product, issuing of a new bond...
- We postulate the event dates are known and we are interested in verifying that the events do not alter the relation between explanatory and dependent variables
- This lead us to formalize the test for structural breaks, or **Chow** test, a test that is equivalent to a test for restrictions on the parameters of a linear model

Diagnostic analysis

- Let us assume that we have data from $t = 1$ to $t = T$ and that the break date is known and equal to $t = m$
- The most general case refer to the two equations

$$\begin{cases} z_t = \alpha_1 + \beta_1 z_{M,t} + \varepsilon & t = 1, 2, \dots m \\ z_t = \alpha_2 + \beta_2 z_{M,t} + \varepsilon & t = m+1, m+2, \dots T \end{cases}$$

- The null hypothesis of the Chow test postulates that the break do not alter the relation across variables, that it $H_0 : \alpha_1 = \alpha_2 \cup \beta_1 = \beta_2$
- Under the null, the model becomes

$$z_t = \alpha + \beta z_{M,t} + \varepsilon \quad t = 1, 2, \dots t$$

Diagnostic analysis

- For the construction of the test statistic we estimate the most general model: we estimate the two equations in the two sample separately, obtaining RSS_1 and RSS_2
- We then estimate the model in the full sample, obtaining RSS_T
- Under the null hypothesis, the parameter are stable in the two sub-samples, while under the break, the parameters differ
- The most general model has twice the parameters of the simplest (restricted) model, leading to a test statistic equivalent to that for parameters restrictions (we impose k restrictions)

$$\frac{\left(RSS_T - RSS_1 - RSS_2\right)/(k)}{RSS_1 + RSS_2/\left(T - 2k\right)} \sim F_{k,T-2k}$$

Diagnostic analysis

- The Chow test is a linear restriction test as we might write the model as

$$\begin{bmatrix} \boldsymbol{z}_1 \\ \boldsymbol{z}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_m \alpha_1 \\ \mathbf{1}_{T-m} \alpha_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{z}_{M,1} & 0 \\ 0 & \boldsymbol{z}_{M_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$$

  where we collect sub-sample data for both the dependent and the explanatory
- Note that this model has parameters specific to the sub-samples, but all parameters can be estimated in a single step by OLS
- The no-break case is a restricted version of the system, thus making appropriate the evaluation with a standard test for linear restrictions (in this case equality restrictions) among parameters

Diagnostic analysis

- But what happens if we postulate the existence of a break but we do not know the break date?
- One of the most common approaches is to run the Chow test for different potential break dates...
- Assume that the total sample is T, then set $\tau = 10\%$ of the data and $\tau T = p$
- Run the Chow test for all possible break dates $m = p, p+1, p+2, \ldots T-p$ and store the test statistic (or the p-value)
- Plot the test statistic (p-value) over time and identify the maximum (minimum) of the test statistic (p-value); the date of the maximum (minimum) is the candidate break date
- Of course, looking at the date is relevant only if the test statistic (p-value) is above (below) the critical value (confidence level)

Diagnostic analysis

- We now move to hypotheses regarding the innovations ε_t
- We did not make an assumption about the errors distribution, so we will not proceed to evaluate distributional hypotheses; the statistical literature includes several proposals, and among them I just cite the Jarque-Bera test for normality and Kolmogorov-Smirnov test for distributional hypothesis
- Back to the hypotheses of the Gauss-Markov theorem: homoskedasticity of residuals; residuals serially uncorrelated
- Let's first focus on heteroskedasticity, the occurrence of variances changing over subjects (for cross-sectional dataset) or over time (for time series datasets); we will focus on one possible way of addressing heteroskedasticity, in the Quantitative Risk Management course you will see models designed to capture heteroskedasticity with variances changing over time

Diagnostic analysis

- Remember that the assumptions behind Gauss-Markov allow us to define the OLS as the best linear unbiased estimator
- In the absence of some of those hypotheses, this interpretation of the OLS is no more valid
- In addition, we will see which are the consequences of heteroskedasticity in the evaluation of the error variance and, consequently, in performing inference on the model parameters

Diagnostic analysis

- The two forms of heteroskedasticity can be identified by means of proper statistical tools
- In the case innovation variances changing across subjects, that is when we consider cross-sectional datasets, the most common test is the White test
- Consider the regression

$$y_i = \alpha + \beta x_i + \delta z_i + \varepsilon_i$$

- The Gauss-Markov theorem requires that $\mathbb{V}[\varepsilon_i] = \sigma^2$, and under heteroskedasticity we have $\mathbb{V}[\varepsilon_i] = \sigma_i^2$
- To detect heteroskedasticity we take the residuals of the fitted model $\hat{\varepsilon}_i$ and use them to build an auxiliary regression

$$\hat{\varepsilon}_i^2 = \omega + \gamma_1 x_i + \gamma_2 z_i + \gamma_3 x_i^2 + \gamma_4 z_i^2 + \gamma_5 x_i z_i + \eta_i$$

that is, we regress squared residuals on covariates, their squares and their cross-products

Diagnostic analysis

- We do have heteroskedasticity if the coefficients of covariates, squared covariates and cross-products are not null
- The null hypothesis become $H_0 : \gamma_1 = \gamma_2 = \ldots \gamma_5 = 0$
- We do not include the intercept ω in the restrictions; under the null hypothesis ω is the variance of the errors
- The test is (again) a test for linear restrictions on parameters, in this case a test of significance of the entire regression
- The test might be run also with time series data, where we are postulating that some of the explanatory variables (or squared values or cross-products) might be related to residuals variance, thus making the variance changing over time
- We might use the usual F-statistic or focus on the asymptotic version, that is $nR^2 \sim \chi^2_q$ with n being the sample size (T with time series data) and q the number of parameters in the auxiliary regression (intercept excluded)

Diagnostic analysis

- Next, we consider another hypotheses made on the error terms, the *independence*, which can be formalized, in a time series setting, as

$$\mathbb{E}\left[\varepsilon_t \varepsilon_{t-k}\right] = 0, \quad k > 0$$

- The *dependence* between innovations is also defined *serial correlation* when we focus on linear regressions for time series
- *Dependence* implies that error terms are *correlated* across time, and is associated with past values having an impact on present (and future) values
- We will consider two tests designed to detect serial correlation in the innovations (a third test will be discussed when dealing with Time Series Analysis)
- Thus we will interpret these tests as diagnostic tests; serial correlation might be a consequence of model misspecification due to omission of explanatory variables, or due to the omission of dynamic relations between variables

Diagnostic analysis

- The simplest form of serial correlation in the errors is obtained as follows

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

with the errors satisfying

$$\varepsilon_t = \phi \varepsilon_{t-1} + u_t$$

- The errors follow a so-called Auto Regressive process of order 1, AR(1) (we will discuss the properties of the model when dealing with Time Series Analysis)
- In the model above, the innovations at time t are correlated with (they are a function of) innovations at time $t-1$, thus breaking the independence hypothesis

Diagnostic analysis

- Independence is associated with a null value of coefficient ϕ
- A standard test for verifying the null hypothesis $H_0 : \phi = 0$ is the Durbin-Watson test
- This test is very common, available in the default output of most software performing linear regression
- The construction of the test statistic requires that the estimated model *includes* the intercept and *does not include* the lagged dependent variable among the explanatory variables
- The alternative hypothesis of the test is that $H_1 : \phi > 0$ and the test statistic is based on the estimated residuals

Diagnostic analysis

- The Durbin-Watson test statistic is

$$DW = \frac{\sum_{t=2}^{T} \left(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1}\right)^2}{\sum_{t=2}^{T} \hat{\varepsilon}_t^2}$$

- The test statistic value might be interpreted and linked to the value of the unknown parameter ϕ: a value close to 0 signals $\phi > 0$, a value close to 2 suggests $\phi \approx 0$, while a value close to 4 indicates $\phi < 0$
- The distribution of the test statistic is non-standard and the null hypothesis is evaluated by contrasting the test statistic with limiting values which are determined under the null and under the alternative hypotheses
- If the estimated test statistic is above the upper limit d_U we do not reject the null hypothesis, that is we conclude there is no serial correlation in the residuals

Diagnostic analysis

- If the test statistic is below the lower limit d_L, then we reject the null hypothesis and we conclude that $\phi > 0$
- If the test statistic is between d_L and d_U, the test is inconclusive
- The limits d_L and d_U depend on both the sample size and the number of explanatory variables, and they are tabulated
- The distance between the two limits decreased with increasing T
- Note: the Durbin-Watson test is appropriate to detect dependence between ε_t and ε_{t-1} but dependence might involve ε_t and ε_{t-k} with $k > 1$; in this case we need a more general test

Diagnostic analysis

- The test of Breusch-Godfrey is designed to that purpose as it detect dependence between ε_t and ε_{t-k} with k up to a value p
- The null hypothesis is independence and the test statistic is evaluated by means of an auxiliary regression
- Assume you run the regression of interest

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

and you have estimated the innovations $\hat{\varepsilon}_t$; then, consider the auxiliary regression

$$y_t = \alpha + \beta x_t + \delta_1 \hat{\varepsilon}_{t-1} + \delta_2 \hat{\varepsilon}_{t-2} \ldots + \delta_p \hat{\varepsilon}_{t-p} + \eta_t$$

- The test statistic is

$$BG = TR^2 \sim \chi^2_p$$

Diagnostic analysis

- If we do find evidences of heteroskedasticity and/or serial correlation which are the consequences to our estimates of the model parameters? Let's see this with the model in compact matrix form...

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

- The independence and homoskedasticity hypotheses imply $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma \boldsymbol{I}$, that is, all variances are identical and shocks are uncorrelated
- If we do have only heteroskedasticity, we have $\mathbb{V}[\boldsymbol{\varepsilon}] = \mathit{diag}\left(\sigma_1^2, \sigma_2^2, \ldots \sigma_T^2\right)$, variances change over time but shocks are uncorrelated
- If we do have serial dependence, then $\mathbb{V}[\boldsymbol{\varepsilon}] = \boldsymbol{\Omega}$ with the form of $\boldsymbol{\Omega}$ depending on the type of serial correlation and, in general, the diagonal contains equal elements (i.e, we have homoskedasticity)
- Both serial correlation and heteroskedasticity can be present, leading to $\mathbb{V}[\boldsymbol{\varepsilon}] = \boldsymbol{\Omega}$, and now the diagonal elements will be different

Diagnostic analysis

- Under serial correlation and heteroskedasticity, we have

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right) \\ &= \boldsymbol{\beta}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \\ \mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \beta \end{aligned}$$

- Therefore, we have proved the OLS estimator remains *unbiased* and consistent
- However, for the variance, things are different

Diagnostic analysis

- Under serial correlation and heteroskedasticity, the variance becomes

$$\begin{aligned}\mathbb{V}\left[\hat{\boldsymbol{\beta}}\right] &= \mathbb{E}\left[\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)^{2}\right] = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \\ &= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned}$$

- It is possible to verify that the variance of the OLS estimator does not attain its minimum value, with the subsequent loss of *efficiency*
- This implies that inferential procedures will be affected, standard errors tend to be larger than the correct ones, and test statistics for significance tend to be underestimated

Diagnostic analysis

- To solve this issue, we might use a different estimator for the standard errors, an estimator robust to heteroskedasticity and serial correlation in the error term or *HAC* estimator
- This approach has been proposed by Newey and West and is available in many packages under the name HAC or robust standard errors
- In the simplest case with a single explanatory variable, the estimator of the coefficient variance becomes

$$\mathbb{V}\left[\hat{\beta}\right] = \frac{\sum_{t=1}^{T} \hat{\varepsilon}_t^2 \left(x_t - \bar{x}\right)^2 + 2\sum_{l=1}^{L}\sum_{t=l+1}^{T} \omega_l \hat{\varepsilon}_t \hat{\varepsilon}_{t-l} \left(x_t - \bar{x}\right)\left(x_{t-l} - \bar{x}\right)}{\sum_{t=1}^{T} \left(x_t - \bar{x}\right)^2}$$

where $\omega_l = 1 - \frac{l}{1+L}$ and L might be chosen according to several automatic methods (or calibrated a-priori)

Diagnostic analysis

- Back to our CAPM example with Microsoft and the S&P500 index: is the linear model correctly specified? Let's focus on diagnostic checking...
- **Collinearity:** in the case of the standard CAPM, this is not an issue as the model has a single explanatory variable; when multiple risk factors are included (as in the example with the Large Cap portfolio), linear correlation among explanatory variables can be used to determine possible collinearity problems; this is not a formal test, but large correlation, with values above 0.9, would suggest further checks

	MKT	SMB	HML	RMW	CMA	MOM
MKT	1.	-0.02	-0.15	0.67	-0.02	-0.76
SMB	-0.02	1.	-0.64	0.26	0.69	0.30
HML	-0.15	-0.64	1.	-0.67	-0.98	-0.36
RMW	0.67	0.26	-0.67	1.	0.51	-0.27
CMA	-0.02	0.69	-0.98	0.51	1.	0.51
MOM	-0.76	0.30	-0.36	-0.27	0.51	1.

Diagnostic analysis

- **RESET** test for linearity: estimate the auxiliary regression and compute the F-test statistic

```
X = np.column_stack((np.ones_like(rMKTe), rMKTe))
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit()
Res1.summary()
fit1=Res1.fittedvalues

# RESET test
# Auxiliary regression
X = np.column_stack((np.ones_like(rMKTe[1:n]), rMKTe[1:n], np.
    power(fit1,2), np.power(fit1,3)))
Res2 = sm.OLS(rMSFTe[1:n], X).fit()
# test
RSSR=Res1.ssr
RSSU=Res2.ssr
Fstat=((RSSR-RSSU)/2)/(RSSU/(423))
Pval=1-sp.stats.f.cdf(Fstat,2,423)
Pval
```

- The test has a P-value equal to 0.151 so we do not reject the null and the linear model is appropriate

Diagnostic analysis

- **CHOW** test: two cases, single sample and rolling evaluation

```
# regressions
n=np.size(rMKTe)
m=200
X1 = np.column_stack((np.ones_like(rMKTe[1:m]), rMKTe[1:m]))
X2 = np.column_stack((np.ones_like(rMKTe[m+1:n]), rMKTe[m+1:n]))
Res2a = sm.OLS(rMSFTe[1:m], X1).fit()
Res2b = sm.OLS(rMSFTe[m+1:n], X2).fit()
X = np.column_stack((np.ones_like(rMKTe), rMKTe))
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit()
# Recover RSS
RSSU=Res2a.ssr+Res2b.ssr
RSSR=Res1.ssr
# Build test
Fstat=((RSSR-RSSU)/2)/(RSSU/(n-4))
Pval=1-sp.stats.f.cdf(Fstat,2,n-4)
Pval
```

- The p-value is now 0.02, thus we reject the null at the 5% confidence level

Diagnostic analysis

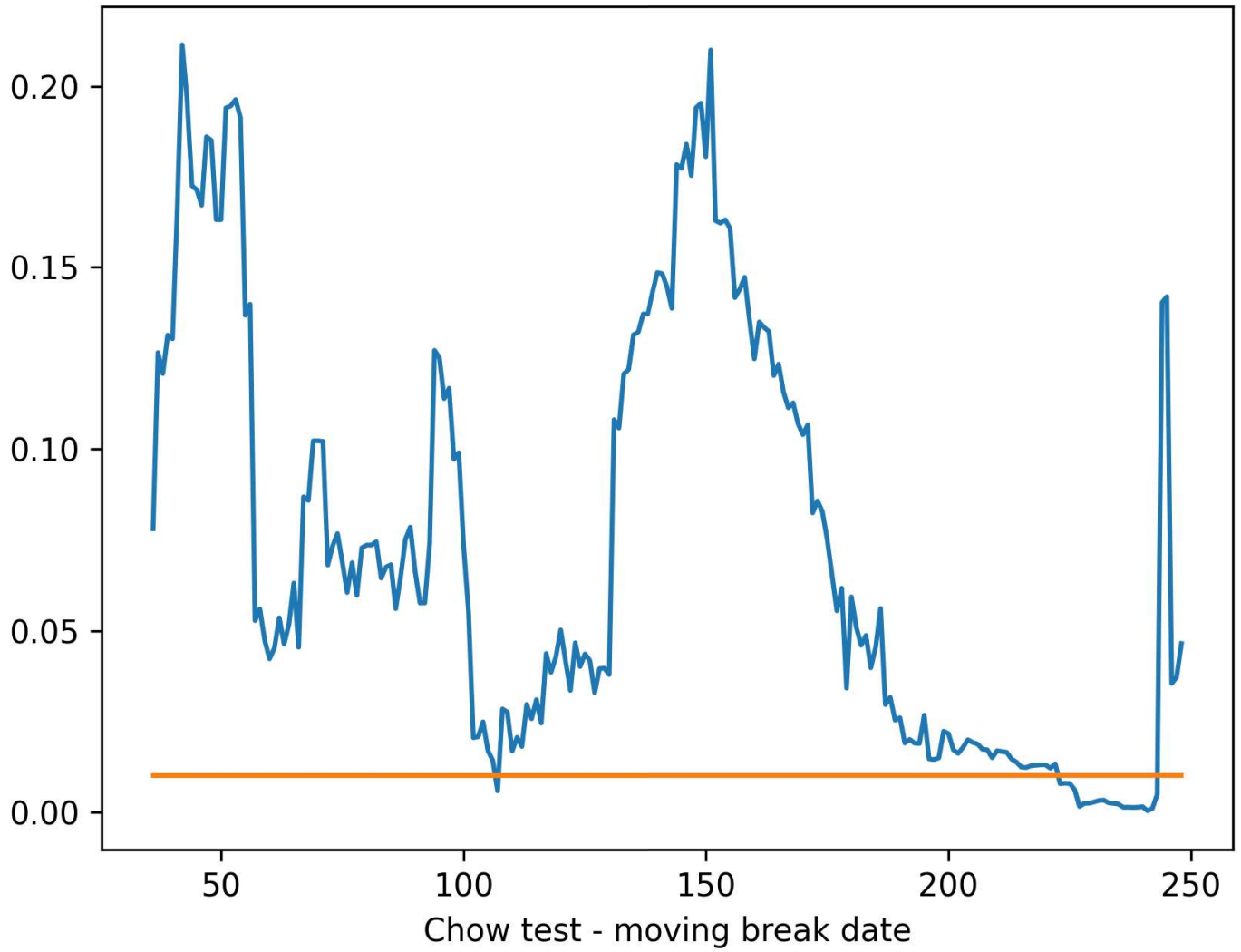
- Chow test with looping across break dates and plot...

```
# looping across break dates
X = np.column_stack((np.ones_like(rMKTe), rMKTe))
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit()
RSSR=Res1.ssr
w=36
i1=w
i2=n-w
Fstat=np.empty(n-w-w+1, dtype=float)
Pval=np.empty(n-w-w+1, dtype=float)
for ii in range(i1,i2):
  X1 = np.column_stack((np.ones_like(rMKTe[1:ii]), rMKTe[1:ii]))
  X2 = np.column_stack((np.ones_like(rMKTe[ii+1:n]), rMKTe[ii+1:
       n]))
  Res2a = sm.OLS(rMSFTe[1:ii], X1).fit()
  Res2b = sm.OLS(rMSFTe[ii+1:n], X2).fit()
  RSSU=Res2a.ssr+Res2b.ssr
  Fstat[ii-w+1]=((RSSR-RSSU)/2)/(RSSU/(n-4))
  Pval[ii-w+1]=1-sp.stats.f.cdf(Fstat[ii-w+1],2,n-4)

plt.plot(range(i1,i2+1),Pval,range(i1,i2+1),0.01*np.ones_like(
    Pval))
plt.xlabel('Chow test - moving break date')
plt.savefig('chowmoving.png',dpi=300)
```

Diagnostic analysis

Moving evaluation of the P-value for the Chow test



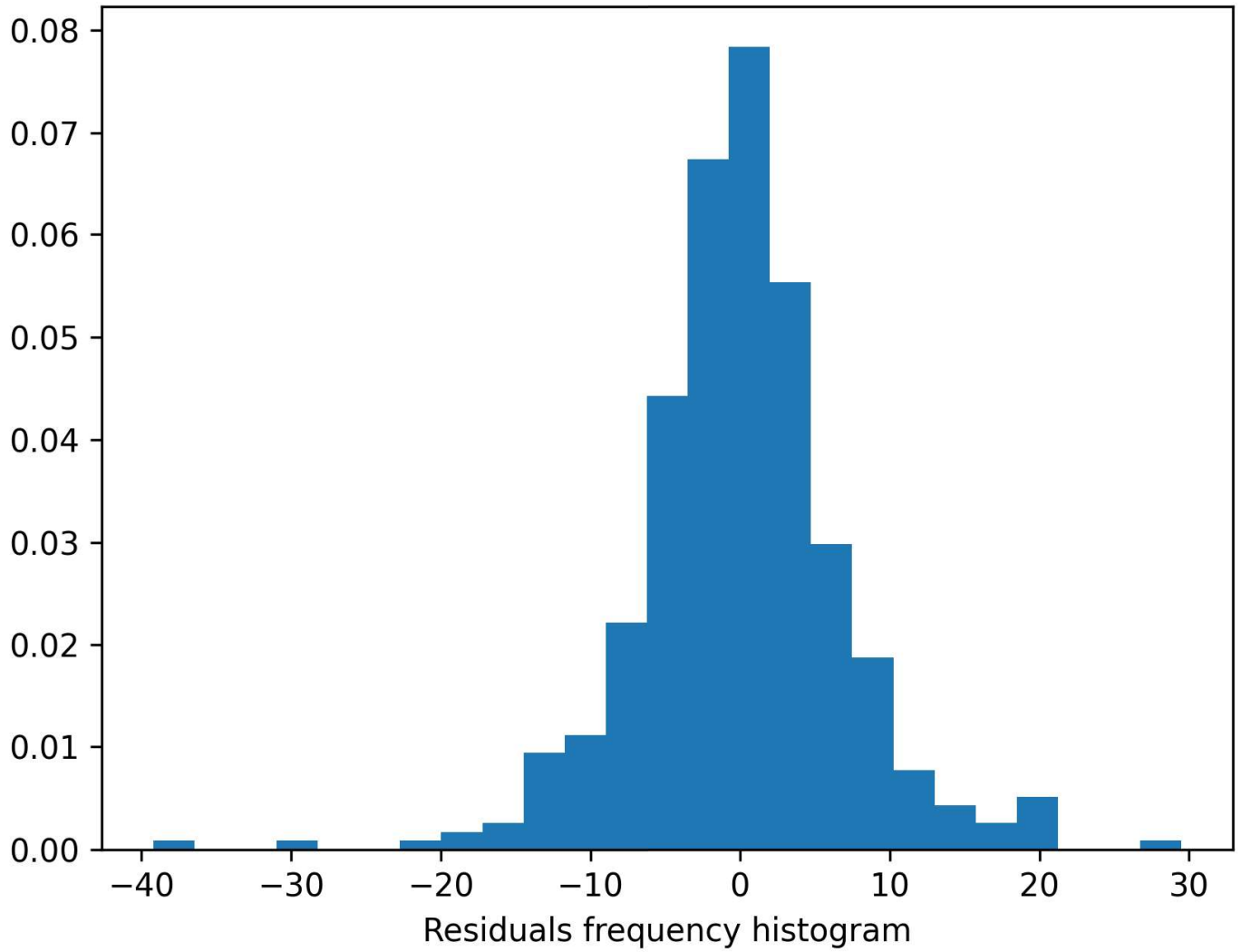
Diagnostic analysis

- Move to the evaluation of residuals, with frequency histogram and probability plot

```
# Residuals plot
X = np.column_stack((np.ones_like(rMKTe), rMKTe))
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit()
rs = Res1.resid
# frequency histogram
plt.hist(rs,bins=25,density=True)
plt.xlabel('Residuals frequency histogram')
plt.savefig('reshist.png',dpi=300)
# probability plot
sp.stats.probplot(rs, plot=plt)
plt.savefig('resprplot.png',dpi=300)
```

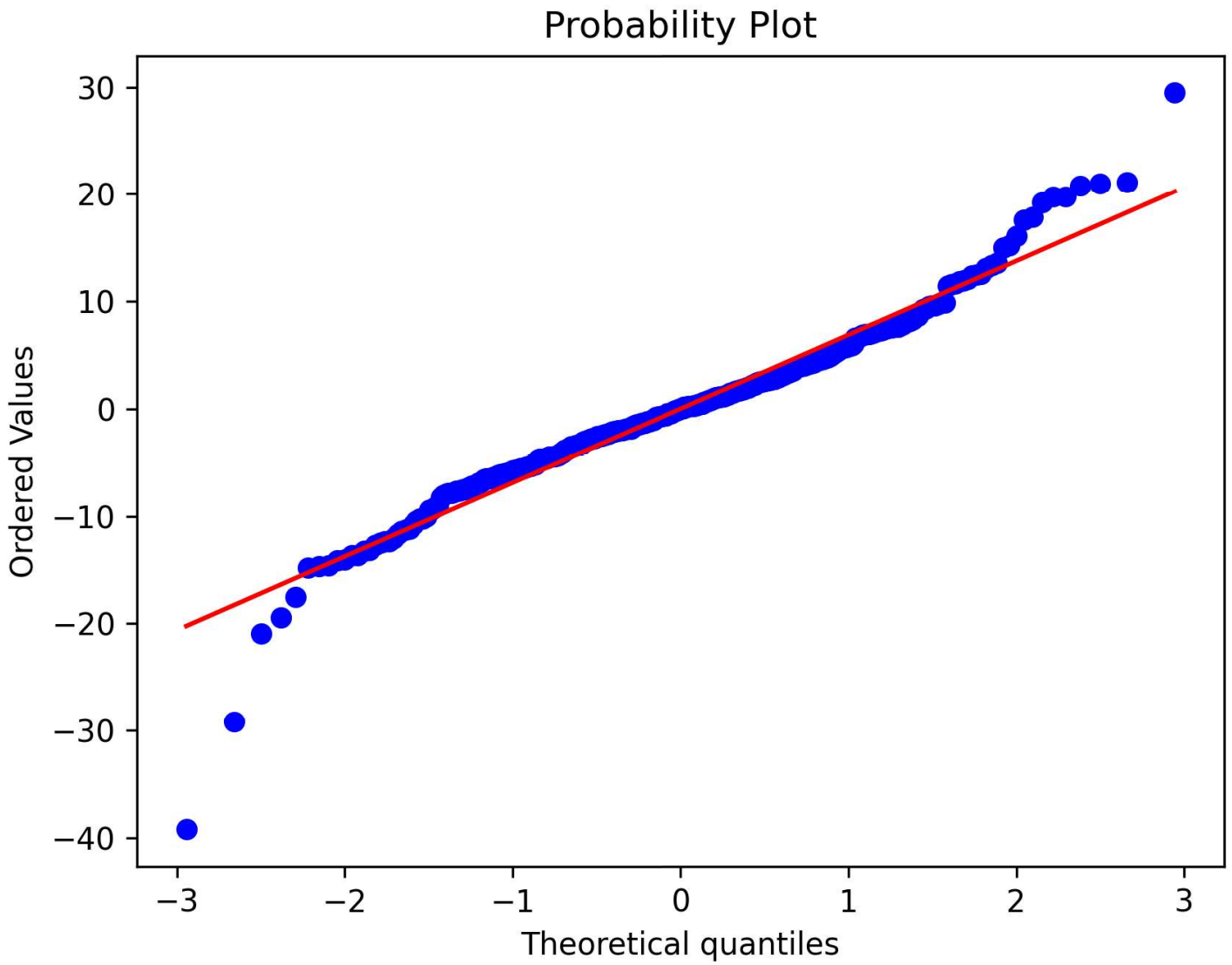
Diagnostic analysis

Frequency histogram



Diagnostic analysis

Probability plot



Diagnostic analysis

- Consider now other indicators and tests...

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   MSFT   R-squared:                       0.332
Model:                            OLS   Adj. R-squared:                  0.330
Method:                 Least Squares   F-statistic:                     211.3
Date:                Tue, 12 Sep 2023   Prob (F-statistic):           3.82e-39
Time:                        15:36:01   Log-Likelihood:                -1436.9
No. Observations:                 427   AIC:                             2878.
Df Residuals:                     425   BIC:                             2886.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.7438      0.343      2.168      0.031       0.069       1.418
Market         1.1576      0.080     14.536      0.000       1.001       1.314
==============================================================================
Omnibus:                       45.093   Durbin-Watson:                   2.207
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              245.923
Skew:                          -0.182   Prob(JB):                     3.97e-54
Kurtosis:                       6.700   Cond. No.                         4.36
==============================================================================
```

Diagnostic analysis

- The standard OLS output provides several elements:

1. The F-statistic for the significance of the entire regression (joint test for significance of all parameters, constant excluded)
2. Jarque-Bera and Omnibus are normality tests (null hypothesis is normality)
3. Skew and Kurtosis refer to the residuals Skewness and Kurtosis (reference values 0 and 3, respectively, leading to Gaussian residuals)
4. Durbin-Watson is a test for first order serial correlation on the residuals (compare values to tabulated values - for the provided example with 1 explanatory + intercept, $T = 427$, $d_L = 1.65$ and $d_U = 1.69$, a value of 2.2 suggest no serial correlation)
5. Cond. No. refers to the condition number of the matrix $\boldsymbol{X}'\boldsymbol{X}$ (the condition number is the square root of the ratio between the largest and the smallest characteristic roots of the matrix, after scaling the column to unit length - large value are worrying for possible multicollinearity)

Diagnostic analysis

- We now consider the other tests for heteroskedasticity and serial correlation

```
# White test
whitetest=sm.stats.diagnostic.het_white(rs,X[1:n])
whitetest[1]
```

- The White test can be built with an auxiliary regression or the *statsmodels* implementation can be used
- The *het_white* command reports four values: the χ^2_q test, the corresponding p-value, the F test and its p-value

```
# Breusch-Godfrey test
bgtest=sm.stats.diagnostic.acorr_breusch_godfrey(Res1,nlags=3)
bgtest[1]
```

- Similarly, the Breusch-Godfrey test is implemented in *statsmodels*, same output structure of White (Chi-square and F statistics with the the p-values)

Diagnostic analysis

- If heteroskedasticity and/or serial correlation are present in the residuals, robust standard errors must be used
- This is managed by providing the proper input to the OLS command, or using a specific post-estimation command

```
# OLS with HAC standard errors
# Direct estimation
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit(cov_type='HAC',cov_kwds={
    'maxlags':5})
Res1.summary()
# Post estimation
Res1 = sm.OLS(rMSFTe[1:n], X[1:n]).fit()
Res2 = Res1.get_robustcov_results(cov_type='HAC',maxlags=5)
Res2.summary()
```

Diagnostic analysis

OLS Regression Results

Dep. Variable:	y	R-squared:	0.332
Model:	OLS	Adj. R-squared:	0.330
Method:	Least Squares	F-statistic:	176.5
Date:	Fri, 15 Sep 2023	Prob (F-statistic):	6.15e-34
Time:	08:28:37	Log-Likelihood:	-1436.9
No. Observations:	427	AIC:	2878.
Df Residuals:	425	BIC:	2886.
Df Model:	1		
Covariance Type:	HAC		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7438	0.325	2.290	0.022	0.107	1.380
x1	1.1576	0.087	13.287	0.000	0.987	1.328

Omnibus:	45.093	Durbin-Watson:	2.207
Prob(Omnibus):	0.000	Jarque-Bera (JB):	245.923
Skew:	-0.182	Prob(JB):	3.97e-54
Kurtosis:	6.700	Cond. No.	4.36

Notes:
[1] Standard Errors are heteroscedasticity and autocorrelation robust (HAC) using 5 lags and without small sample correction

Diagnostic analysis

- In order to verify if the CAPM holds, two set of hypotheses have to be verified

1. $H_0 : \beta_1 = 0$ - we expect a clear rejection of the null hypothesis for all i
2. $H_0 : \alpha_i = 0$ - this must hold (we do not reject the null hypothesis) for all i

- It is possible to verify that the second hypothesis is crucial, but it cannot be tested on an *equation-by-equation* basis and *system estimation* is needed

Forecasting

- Generally speaking, we might be interested in forecasting the dependent variable according to our *static* linear regression model
- This approach requires the knowledge of the future values for the explanatory variables, which we denote by x_f (a vector of variables at time $f = T + h$ with $h \geq 1$); the x_f might be obtained from satellite models (time series models) or recovered from the scenario forecasts provided by external sources
- The predicted values for y_f correspond to a linear combination of parameters and future values of the explanatory variables

$$\hat{y}_f = \begin{bmatrix} 1 & \mathrm{x}_f' \end{bmatrix} \hat{\boldsymbol{\beta}} = \tilde{x}_f' \hat{\boldsymbol{\beta}}$$

- The point forecast is of interest, but we might also be interested of building a confidence interval which should contain, with a given probability, the true future value of the dependent variable

Forecasting

- The future explanatory variables are assumed to be **knowns** and therefore the forecast is a linear combination of parameters
- Given that the parameters (constant included) are distributed as

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

  it follows that

$$\hat{y}_f = \tilde{x}_f'\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\tilde{x}_f'\boldsymbol{\beta}, \sigma^2 \tilde{x}_f' \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \tilde{x}_f\right)$$

- In addition, for the true future value of the explanatory variables it holds

$$y_f = \tilde{x}_f'\boldsymbol{\beta} + \varepsilon_f$$

Forecasting

- The forecast error is thus equal to

$$e_f = y_f - \hat{y}_f = \tilde{x}_f'\boldsymbol{\beta} + \varepsilon_f - \tilde{x}_f'\hat{\boldsymbol{\beta}} = \varepsilon_f - \tilde{x}_f'\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)$$

- Then, the variance of the forecast error equals

$$\begin{aligned}\mathbb{V}\left[e_f\right] &= \mathbb{V}\left[\varepsilon_f\right] + \tilde{x}_f'\mathbb{V}\left[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right]\tilde{x}_f = \sigma^2 + \sigma^2\tilde{x}_f'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\tilde{x}_f \\ &= \sigma^2\left(1 + \tilde{x}_f'\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\tilde{x}_f\right) = \mathbb{V}\left[y_f - \hat{y}_f\right] = \mathbb{V}\left[\hat{y}_f - y_f\right]\end{aligned}$$

- Note that we have $\mathbb{E}\left[\hat{y}_f\right] = \tilde{x}_f'\boldsymbol{\beta}$, $\mathbb{E}\left[e_f\right] = 0$ and it is possible to verify that

$$\frac{e_f}{\sqrt{\mathbb{V}\left[e_f\right]}} \sim \mathcal{N}\left(0, 1\right)$$

Forecasting

- Finally, a confidence interval for y_f, built upon the forecasts, will be equal to

$$\Phi^{-1}\left(\frac{\alpha}{2}\right) \leq \frac{e_f}{\sqrt{\mathbb{V}\left[e_f\right]}} \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$$

$$\Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\mathbb{V}\left[e_f\right]} \leq e_f \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathbb{V}\left[e_f\right]}$$

$$\Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\mathbb{V}\left[e_f\right]} \leq \hat{y}_f - y_f \leq \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathbb{V}\left[e_f\right]}$$

$$\hat{y}_f - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\mathbb{V}\left[e_f\right]} \leq y_f \leq \hat{y}_f - \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\mathbb{V}\left[e_f\right]}$$

Linear Regression

- [...to be continued if time permits...]

Thanks for the attention!