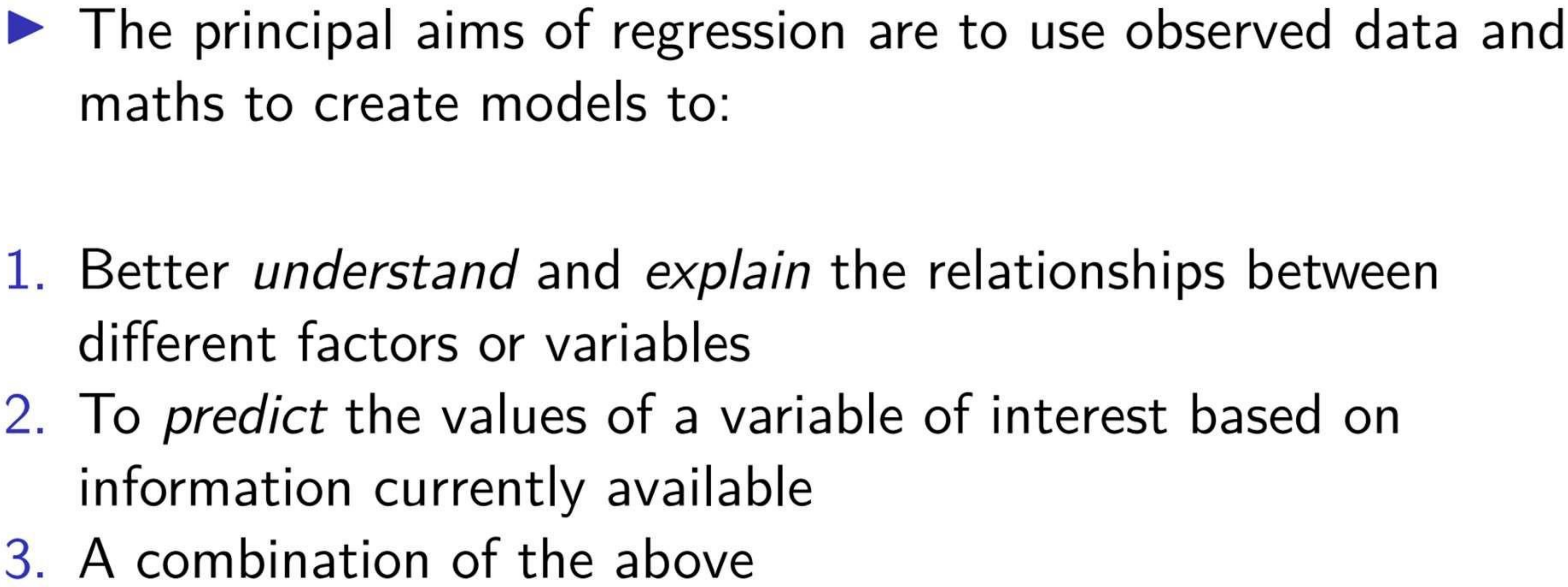
Predictive Analytics For Financial Markets

Kwabena Meneabe Ackon

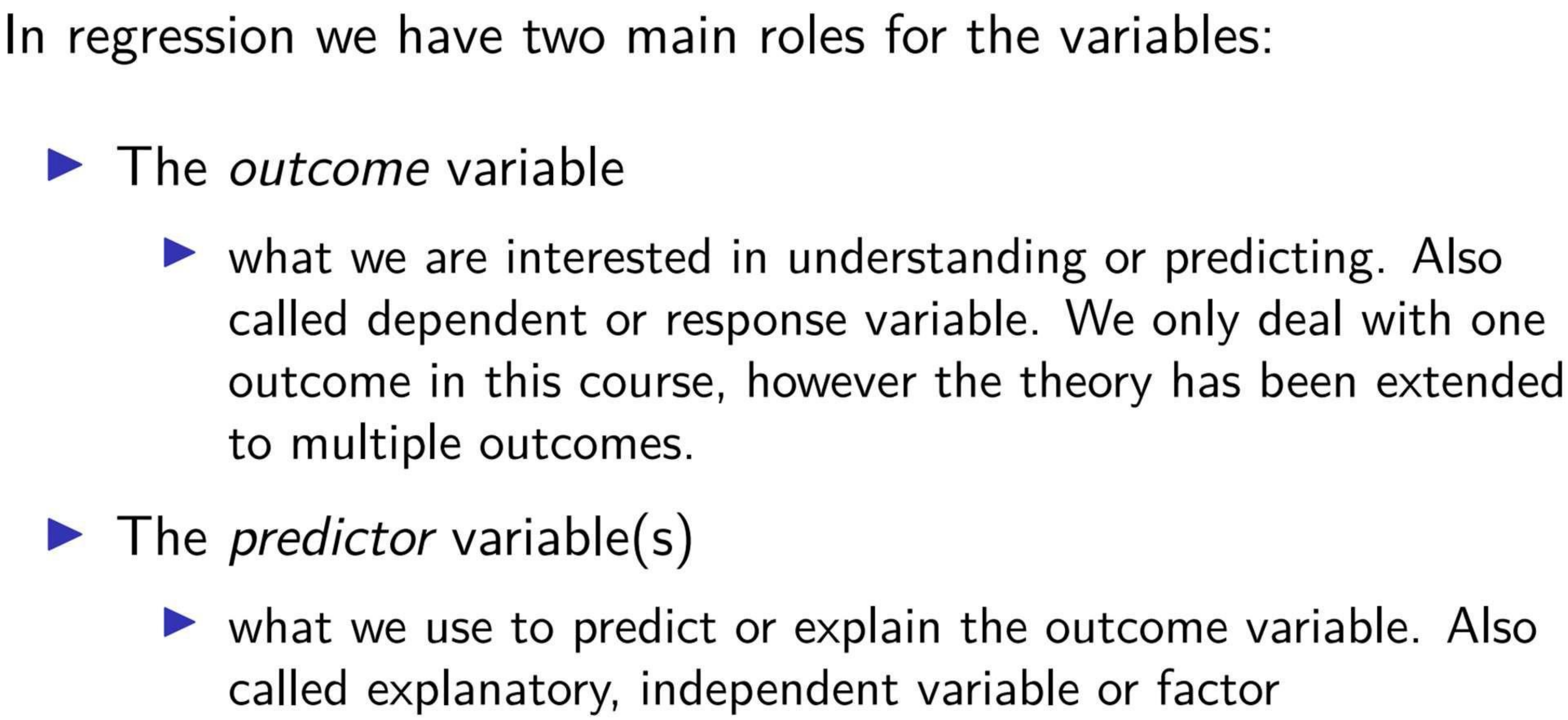
The aims of regression



- The principal aims of regression are to use observed data and maths to create models to:

1. Better *understand* and *explain* the relationships between different factors or variables
2. To *predict* the values of a variable of interest based on information currently available
3. A combination of the above

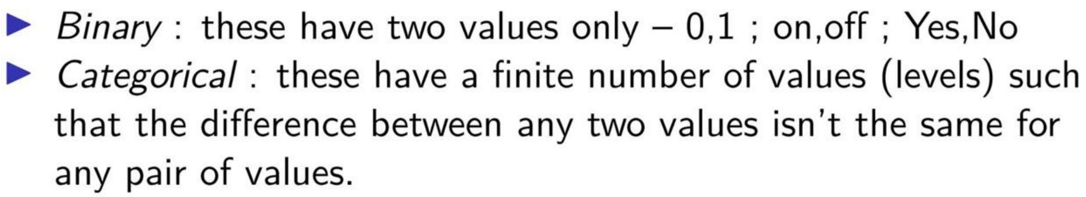
Roles and types of variables



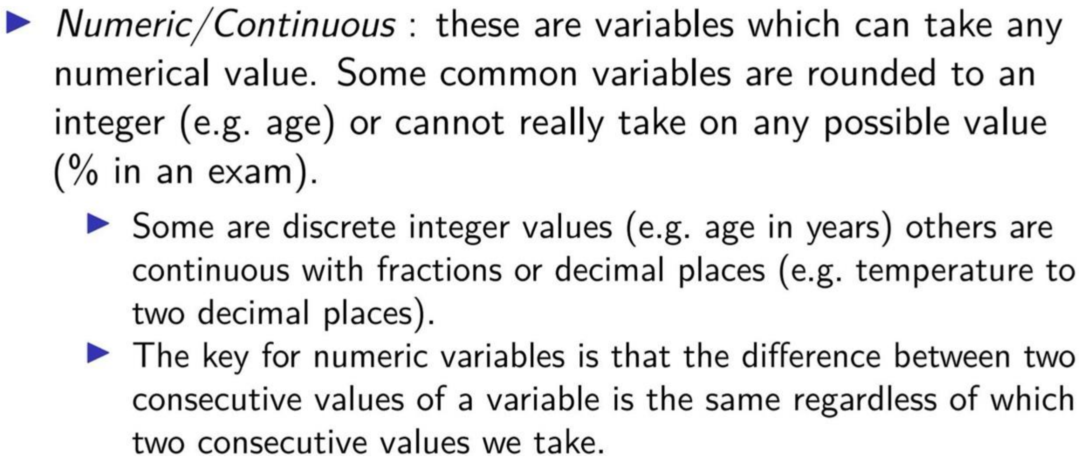
In regression we have two main roles for the variables:

- The *outcome* variable
  - what we are interested in understanding or predicting. Also called dependent or response variable. We only deal with one outcome in this course, however the theory has been extended to multiple outcomes.
- The *predictor* variable(s)
  - what we use to predict or explain the outcome variable. Also called explanatory, independent variable or factor

Type of variable

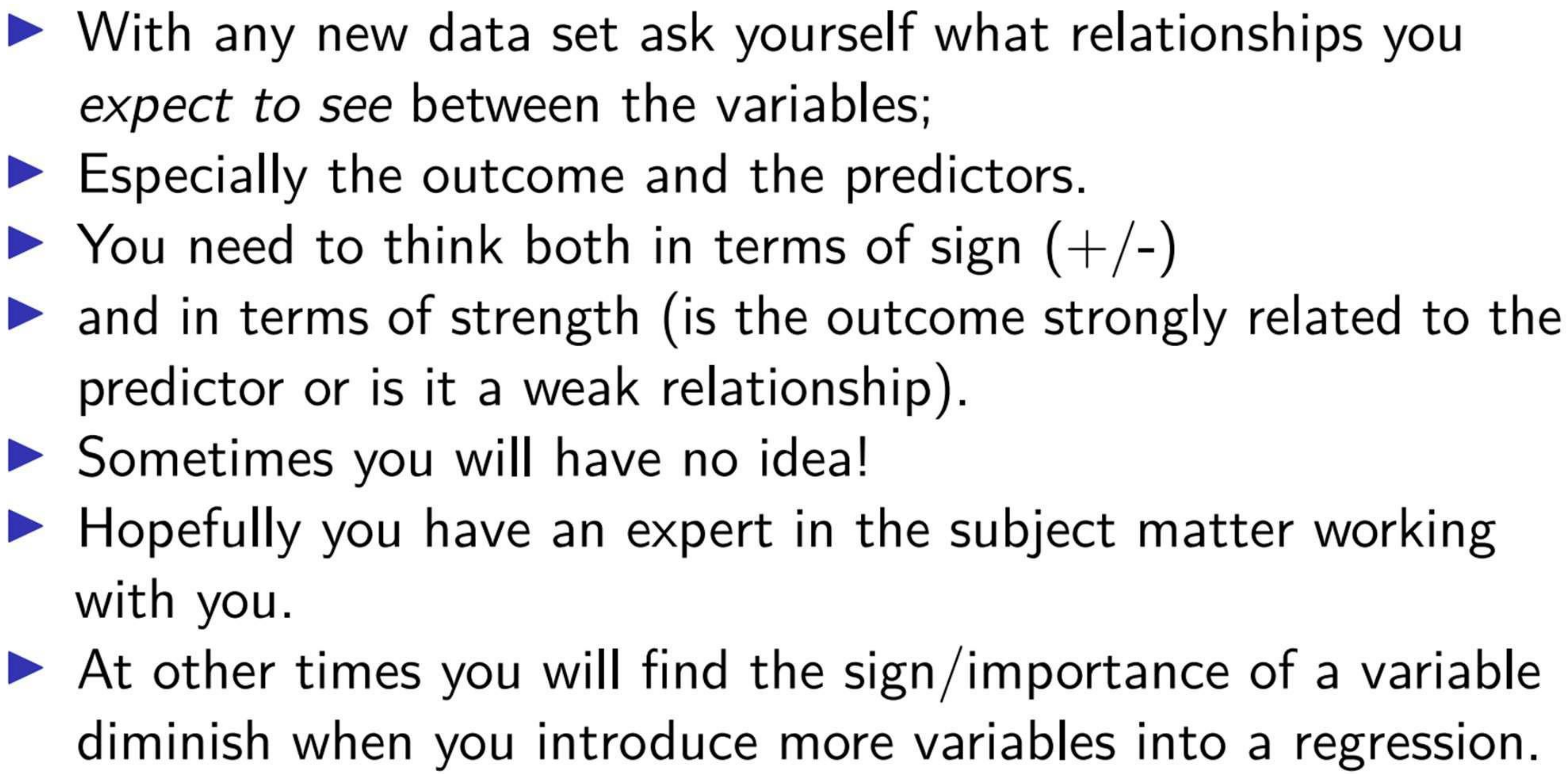


- *Binary* : these have two values only – 0,1 ; on,off ; Yes,No
- *Categorical* : these have a finite number of values (levels) such that the difference between any two values isn't the same for any pair of values.



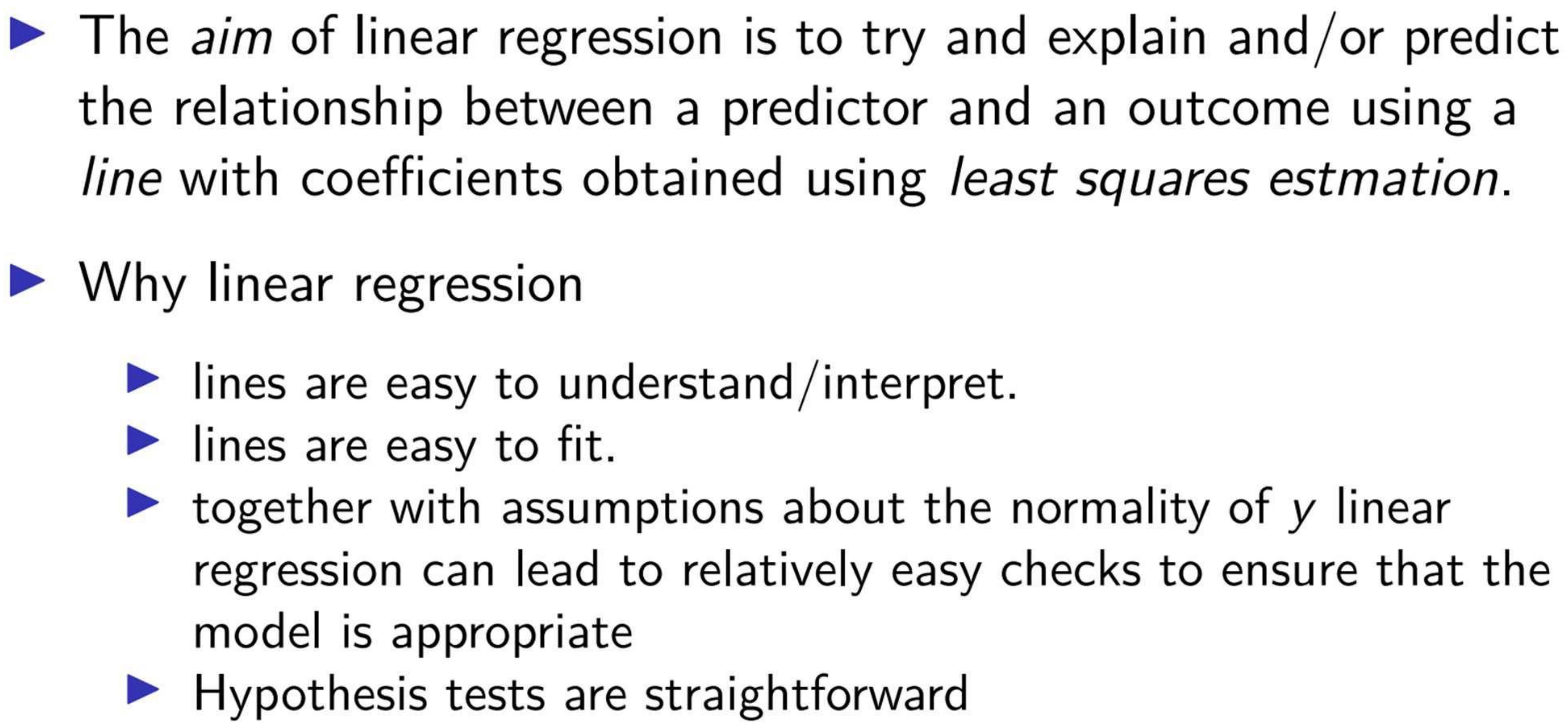
- *Numeric/Continuous* : these are variables which can take any numerical value. Some common variables are rounded to an integer (e.g. age) or cannot really take on any possible value (% in an exam).
  - Some are discrete integer values (e.g. age in years) others are continuous with fractions or decimal places (e.g. temperature to two decimal places).
  - The key for numeric variables is that the difference between two consecutive values of a variable is the same regardless of which two consecutive values we take.

How are the variables related



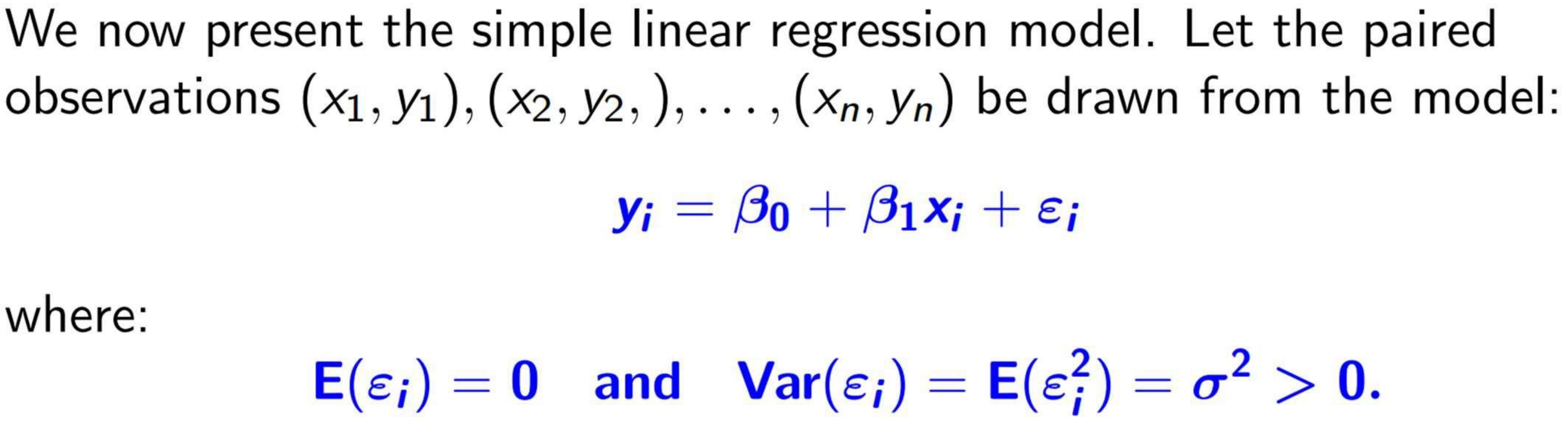
- With any new data set ask yourself what relationships you *expect to see* between the variables;
- Especially the outcome and the predictors.
- You need to think both in terms of sign (+/-)
- and in terms of strength (is the outcome strongly related to the predictor or is it a weak relationship).
- Sometimes you will have no idea!
- Hopefully you have an expert in the subject matter working with you.
- At other times you will find the sign/importance of a variable diminish when you introduce more variables into a regression.

Why linear regression



- The *aim* of linear regression is to try and explain and/or predict the relationship between a predictor and an outcome using a *line* with coefficients obtained using *least squares estmation*.
- Why linear regression
  - lines are easy to understand/interpret.
  - lines are easy to fit.
  - together with assumptions about the normality of y linear regression can lead to relatively easy checks to ensure that the model is appropriate
  - Hypothesis tests are straightforward

Simple linear regression



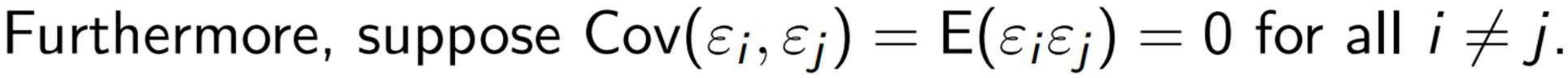
We now present the simple linear regression model. Let the paired observations $(x_1, y_1), (x_2, y_2,), \ldots, (x_n, y_n)$ be drawn from the model:

$$\boldsymbol{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

where:

$$\mathbf{E}(\varepsilon_i) = 0 \quad \textbf{and} \quad \mathbf{Var}(\varepsilon_i) = \mathbf{E}(\varepsilon_i^2) = \sigma^2 > 0.$$

Simple linear regression

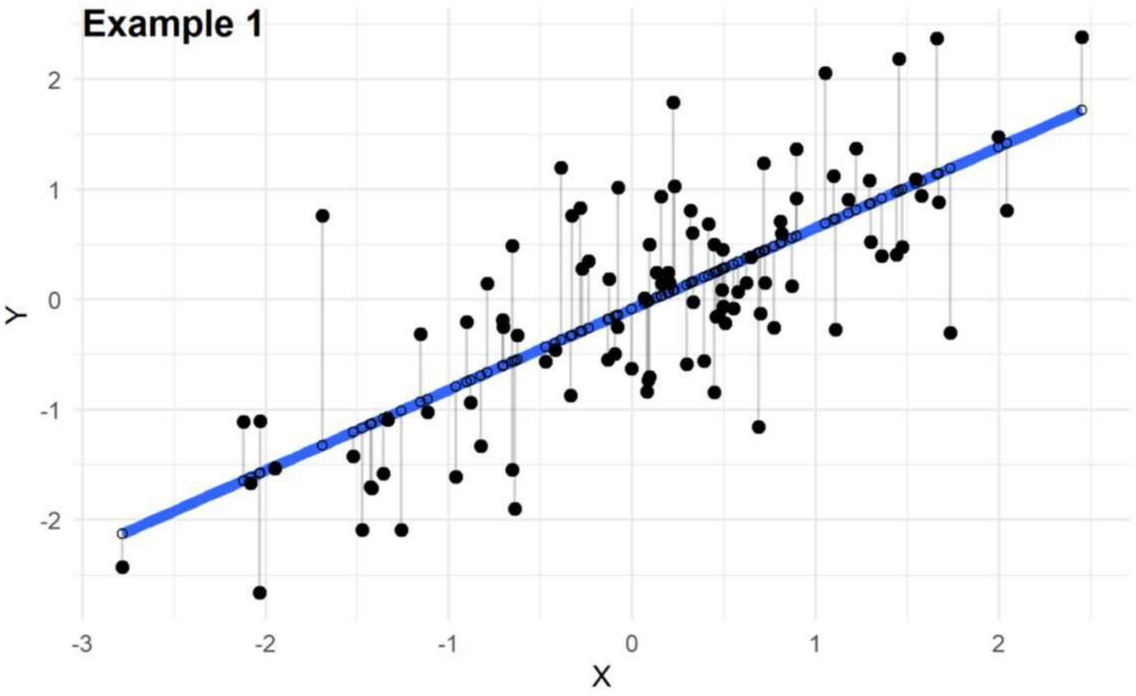


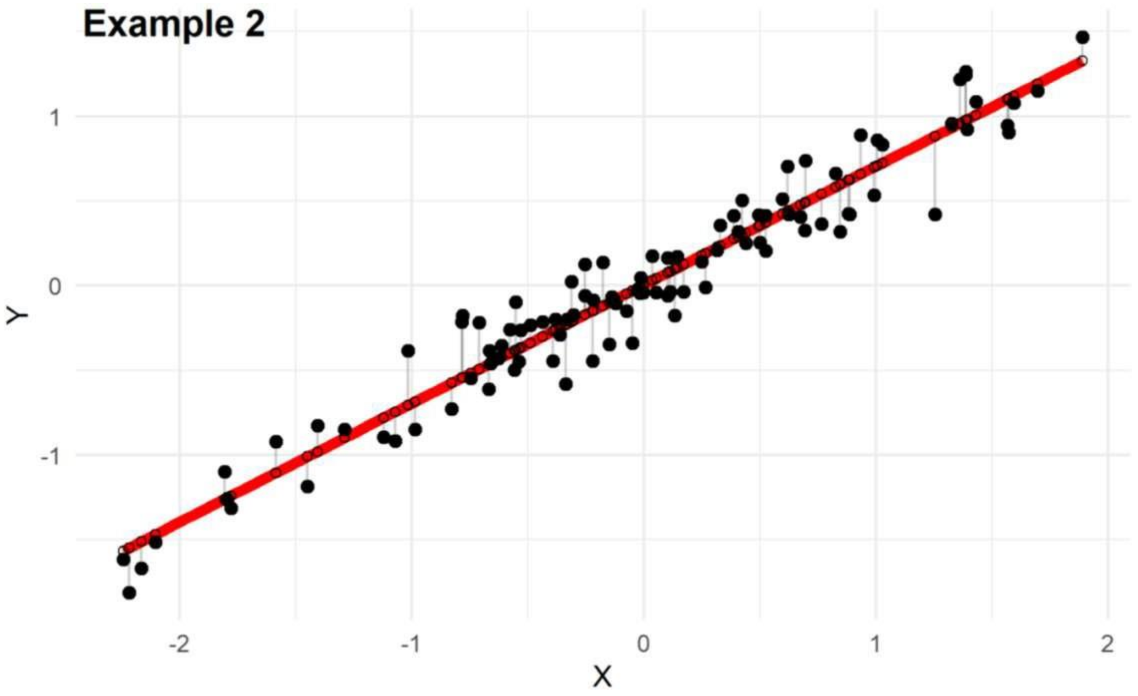
Furthermore, suppose $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = \mathrm{E}(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$.

Sometimes we assume $\varepsilon_i \sim N(0, \sigma^2)$

So the model has three parameters: β_0, β_1 and σ^2.

The residuals





The residuals

- The vertical distance between the line of fit and the observed value is called the *residual*.
- The residuals can be thought of as representing the error in the regression.

$$r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Formally in least squares estimation the quantity that has to be minimised with respect to the parameters β_0 and β_1 is:

$$\begin{aligned} SSE &= \sum r_i^2 \\ &= \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \end{aligned}$$

Capital Asset Pricing Model (CAPM)

The capital asset pricing model (CAPM) is a simple asset pricing model in finance given by:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where y_i is a stock return and x_i is a market return at time i.

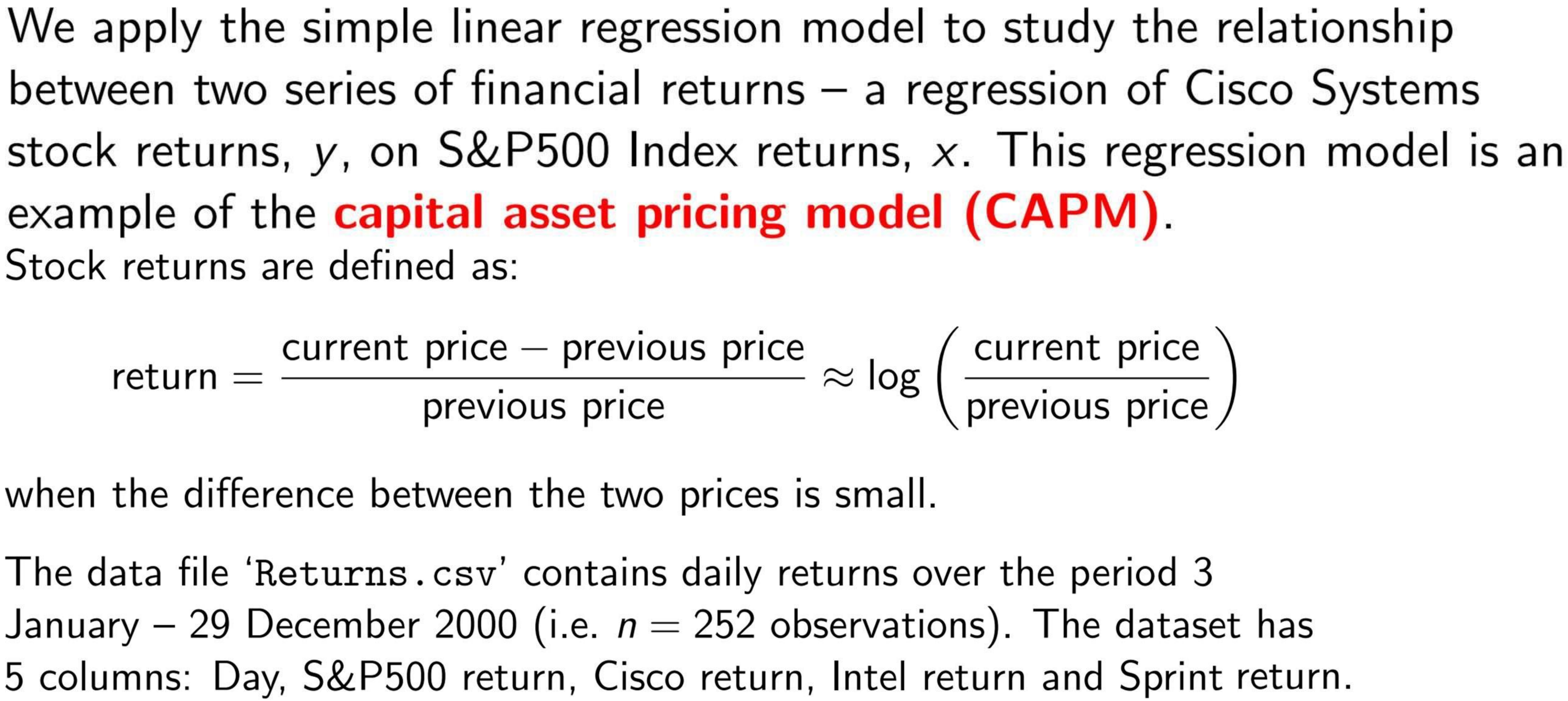
Some remarks are the following.

i. β_1 measures the market-related (or systematic) risk of the stock.

ii. Market-related risk is unavoidable, while firm-specific risk may be 'diversified away' through *hedging*.

iii. Variance is a simple measure (and one of the most frequently-used) of risk in finance.

Capital Asset Pricing Model (CAPM)



We apply the simple linear regression model to study the relationship between two series of financial returns – a regression of Cisco Systems stock returns, y, on S&P500 Index returns, x. This regression model is an example of the **capital asset pricing model (CAPM)**.

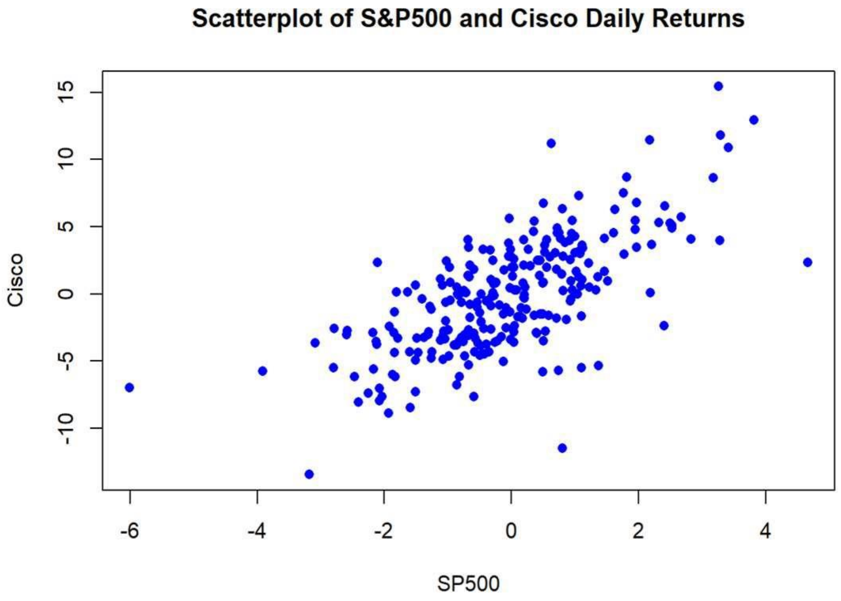
Stock returns are defined as:

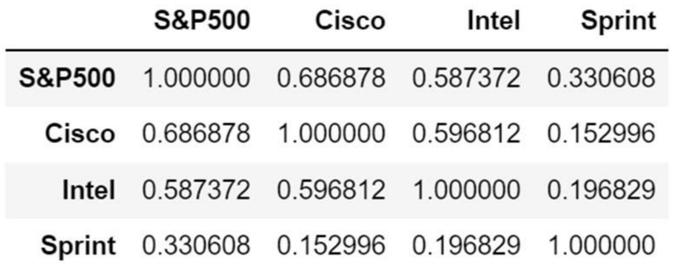
$$\text{return} = \frac{\text{current price} - \text{previous price}}{\text{previous price}} \approx \log\left(\frac{\text{current price}}{\text{previous price}}\right)$$

when the difference between the two prices is small.

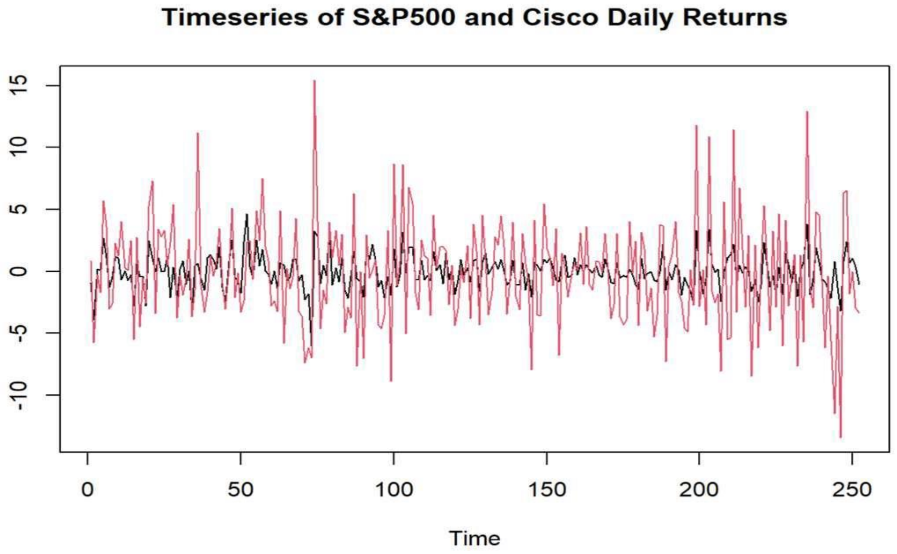
The data file ‘`Returns.csv`’ contains daily returns over the period 3 January – 29 December 2000 (i.e. $n = 252$ observations). The dataset has 5 columns: Day, S&P500 return, Cisco return, Intel return and Sprint return.

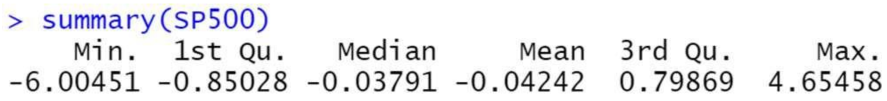
Capital Asset Pricing Model (CAPM)



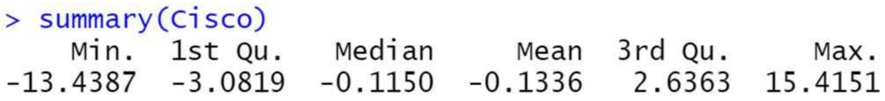


	S&P500	Cisco	Intel	Sprint
S&P500	1.000000	0.686878	0.587372	0.330608
Cisco	0.686878	1.000000	0.596812	0.152996
Intel	0.587372	0.596812	1.000000	0.196829
Sprint	0.330608	0.152996	0.196829	1.000000





```
> summary(SP500)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-6.00451 -0.85028 -0.03791 -0.04242  0.79869  4.65458
```

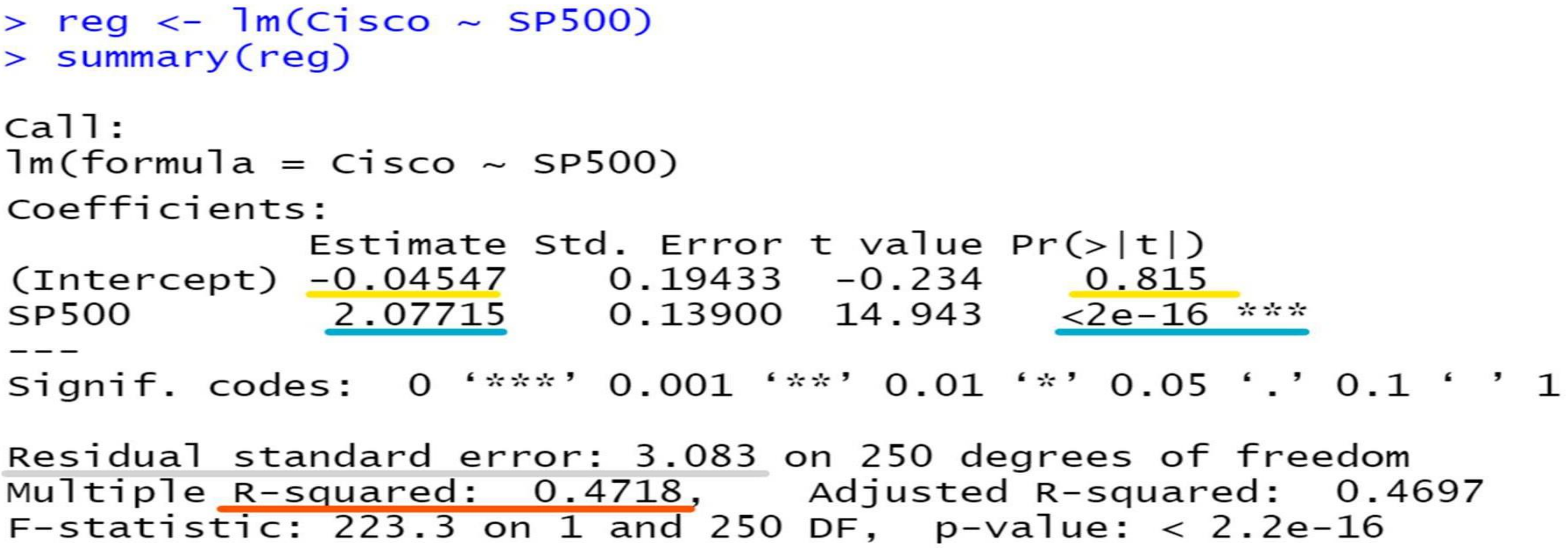


```
> summary(Cisco)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.4387  -3.0819  -0.1150  -0.1336   2.6363  15.4151
```

Capital Asset Pricing Model

We fit the regression model: $\text{Cisco} = \beta_0 + \beta_1 \text{S\&P500} + \varepsilon$.

Our rationale is that part of the fluctuation in Cisco returns was driven by the fluctuation in the S&P500 returns.

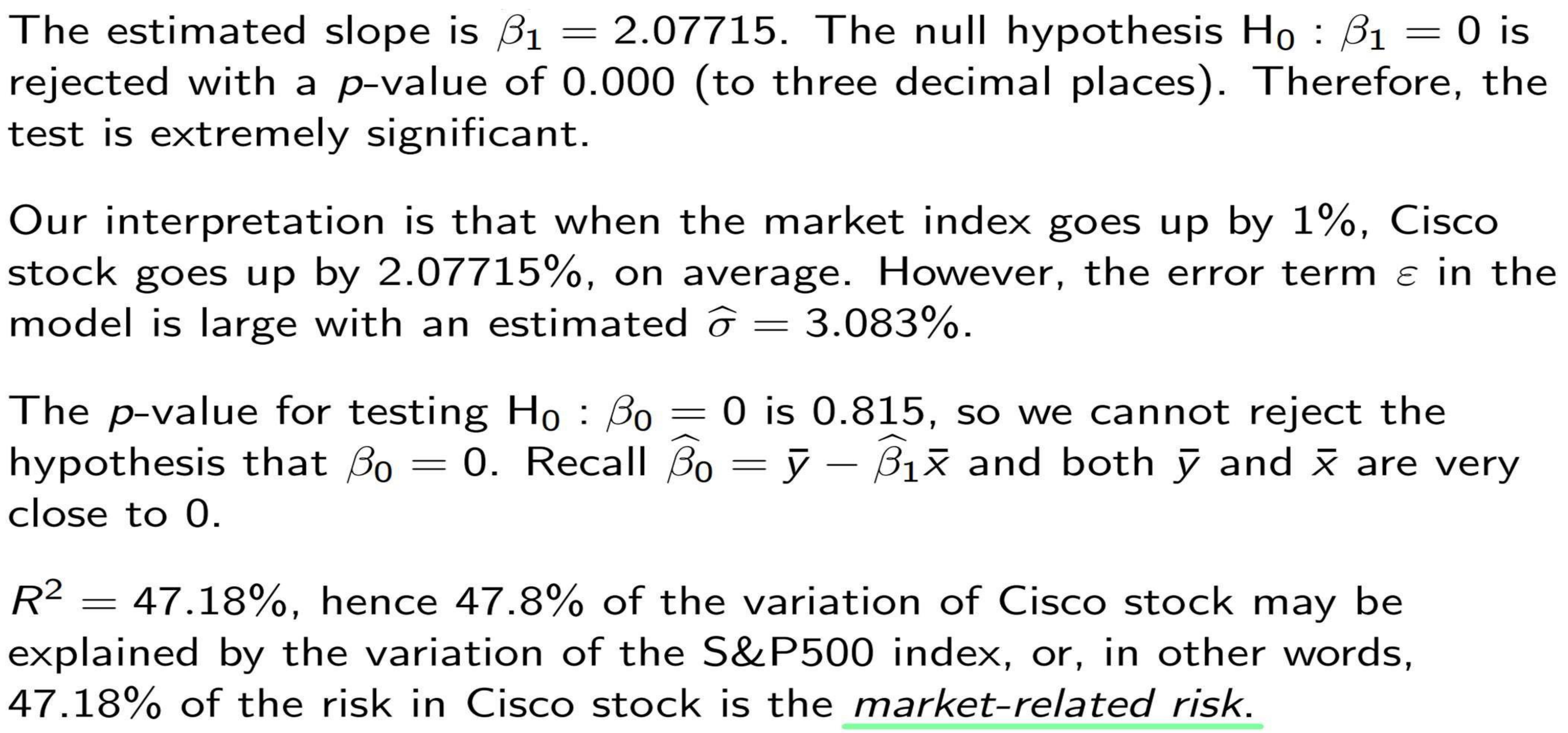


```
> reg <- lm(Cisco ~ SP500)
> summary(reg)

Call:
lm(formula = Cisco ~ SP500)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04547    0.19433  -0.234    0.815
SP500        2.07715    0.13900  14.943   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.083 on 250 degrees of freedom
Multiple R-squared:  0.4718,    Adjusted R-squared:  0.4697
F-statistic: 223.3 on 1 and 250 DF,  p-value: < 2.2e-16
```

Capital Asset Pricing Model (CAPM)



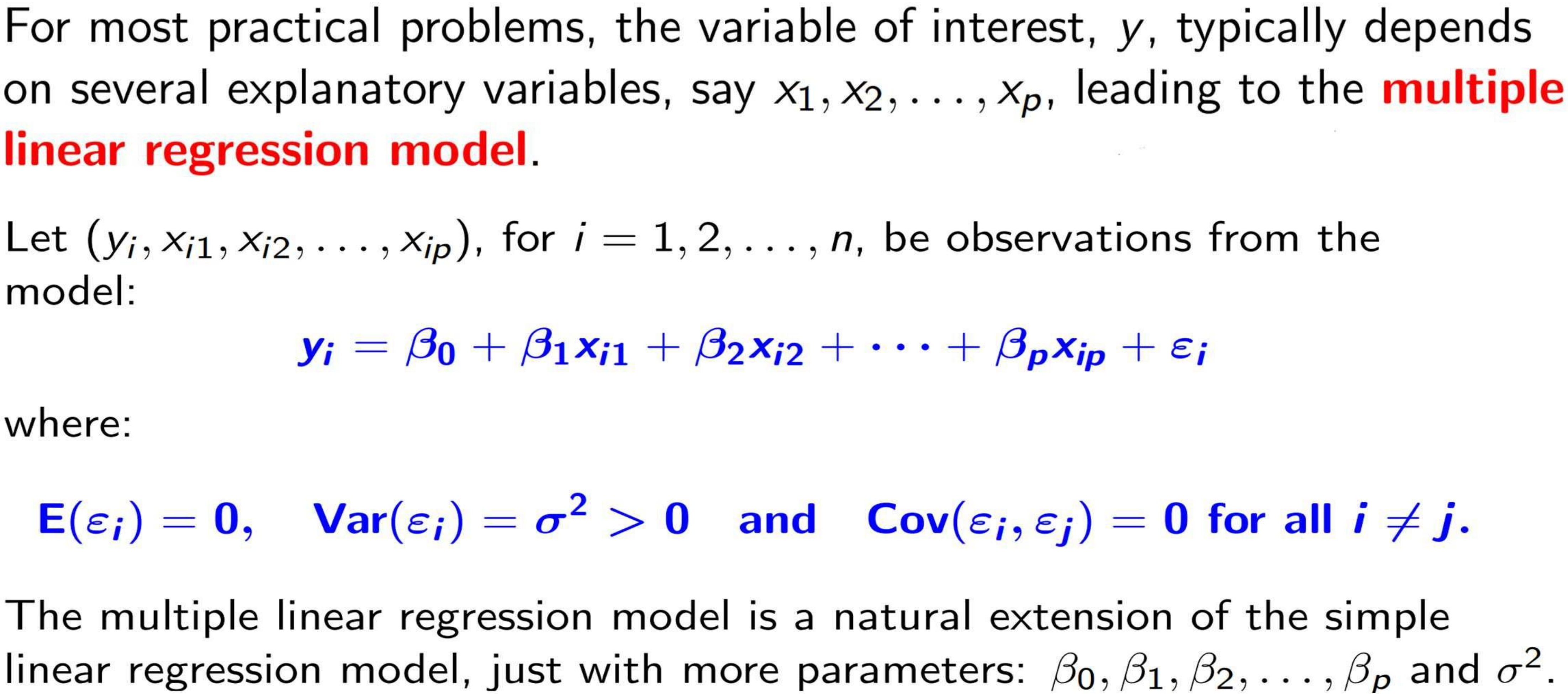
The estimated slope is $\beta_1 = 2.07715$. The null hypothesis $\text{H}_0 : \beta_1 = 0$ is rejected with a p-value of 0.000 (to three decimal places). Therefore, the test is extremely significant.

Our interpretation is that when the market index goes up by 1%, Cisco stock goes up by 2.07715%, on average. However, the error term ε in the model is large with an estimated $\widehat{\sigma} = 3.083\%$.

The p-value for testing $\text{H}_0 : \beta_0 = 0$ is 0.815, so we cannot reject the hypothesis that $\beta_0 = 0$. Recall $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$ and both $\bar{y}$ and $\bar{x}$ are very close to 0.

$R^2 = 47.18\%$, hence 47.8% of the variation of Cisco stock may be explained by the variation of the S&P500 index, or, in other words, 47.18% of the risk in Cisco stock is the *market-related risk*.

Multiple linear regression models



For most practical problems, the variable of interest, y, typically depends on several explanatory variables, say $x_1, x_2, \ldots, x_p$, leading to the **multiple linear regression model**.

Let $(y_i, x_{i1}, x_{i2}, \ldots, x_{ip})$, for $i = 1, 2, \ldots, n$, be observations from the model:

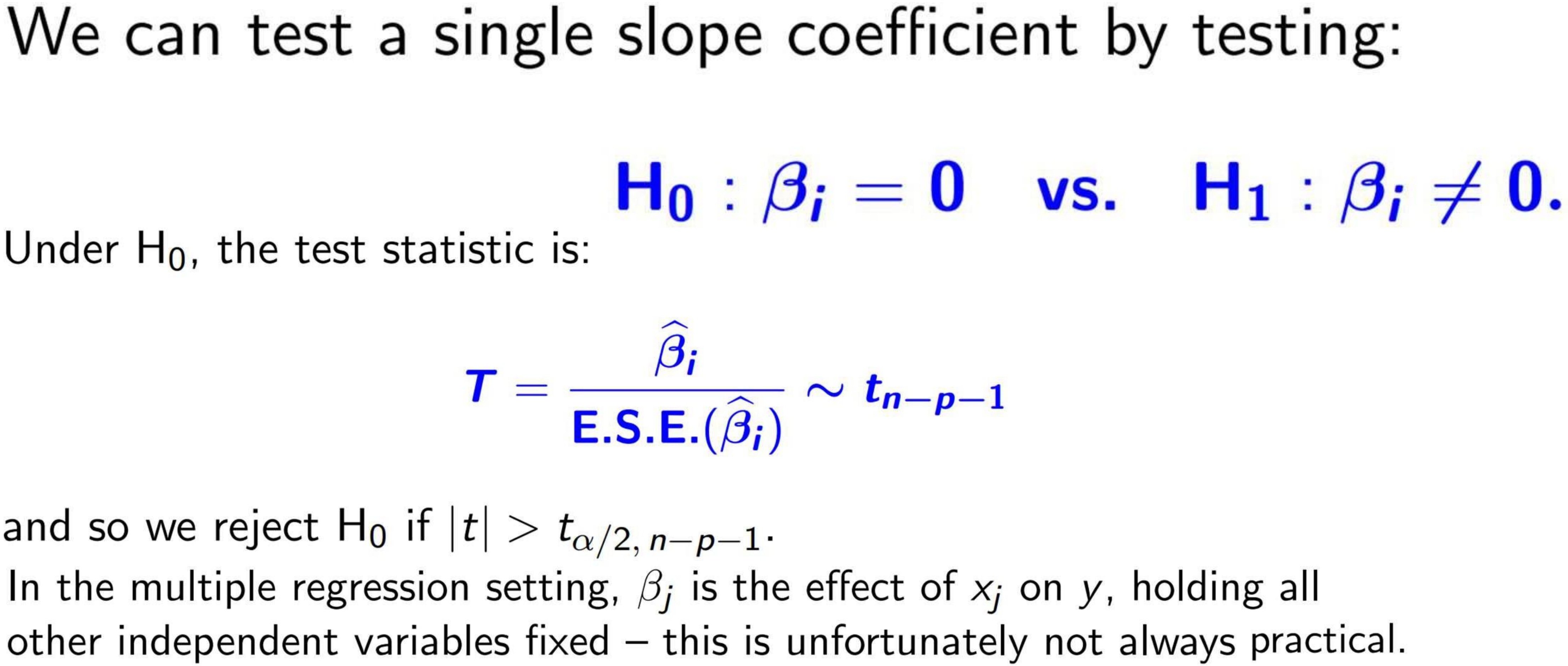
$$\boldsymbol{y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i}$$

where:

$$\mathbf{E}(\boldsymbol{\varepsilon_i}) = \mathbf{0}, \quad \mathbf{Var}(\boldsymbol{\varepsilon_i}) = \boldsymbol{\sigma^2} > \mathbf{0} \quad \textbf{and} \quad \mathbf{Cov}(\boldsymbol{\varepsilon_i}, \boldsymbol{\varepsilon_j}) = \mathbf{0} \textbf{ for all } \boldsymbol{i \neq j}.$$

The multiple linear regression model is a natural extension of the simple linear regression model, just with more parameters: $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and σ^2.

Coefficients and their significance



We can test a single slope coefficient by testing:

$$\mathbf{H_0} : \beta_i = 0 \quad \textbf{vs.} \quad \mathbf{H_1} : \beta_i \neq 0.$$

Under H_0, the test statistic is:

$$T = \frac{\widehat{\beta}_i}{\text{E.S.E.}(\widehat{\beta}_i)} \sim t_{n-p-1}$$

and so we reject H_0 if $|t| > t_{\alpha/2, n-p-1}$.

In the multiple regression setting, β_j is the effect of x_j on y, holding all other independent variables fixed – this is unfortunately not always practical.

F-statistic

It is also possible to test whether all the regression coefficients are equal to zero. This is known as a **joint test of significance** and can be used to test the overall significance of the regression model, i.e. whether there is at least one significant explanatory (independent) variable, by testing:

$$\mathbf{H_0} : \beta_1 = \beta_2 = \cdots = \beta_p = \mathbf{0} \quad \textbf{vs.} \quad \mathbf{H_1} : \textbf{At least one } \beta_i \neq \mathbf{0}.$$

Indeed, it is preferable to perform this joint test of significance *before* conducting t tests of individual slope coefficients.

Failure to reject H_0 would render the model useless and hence the model would not warrant any further statistical investigation.

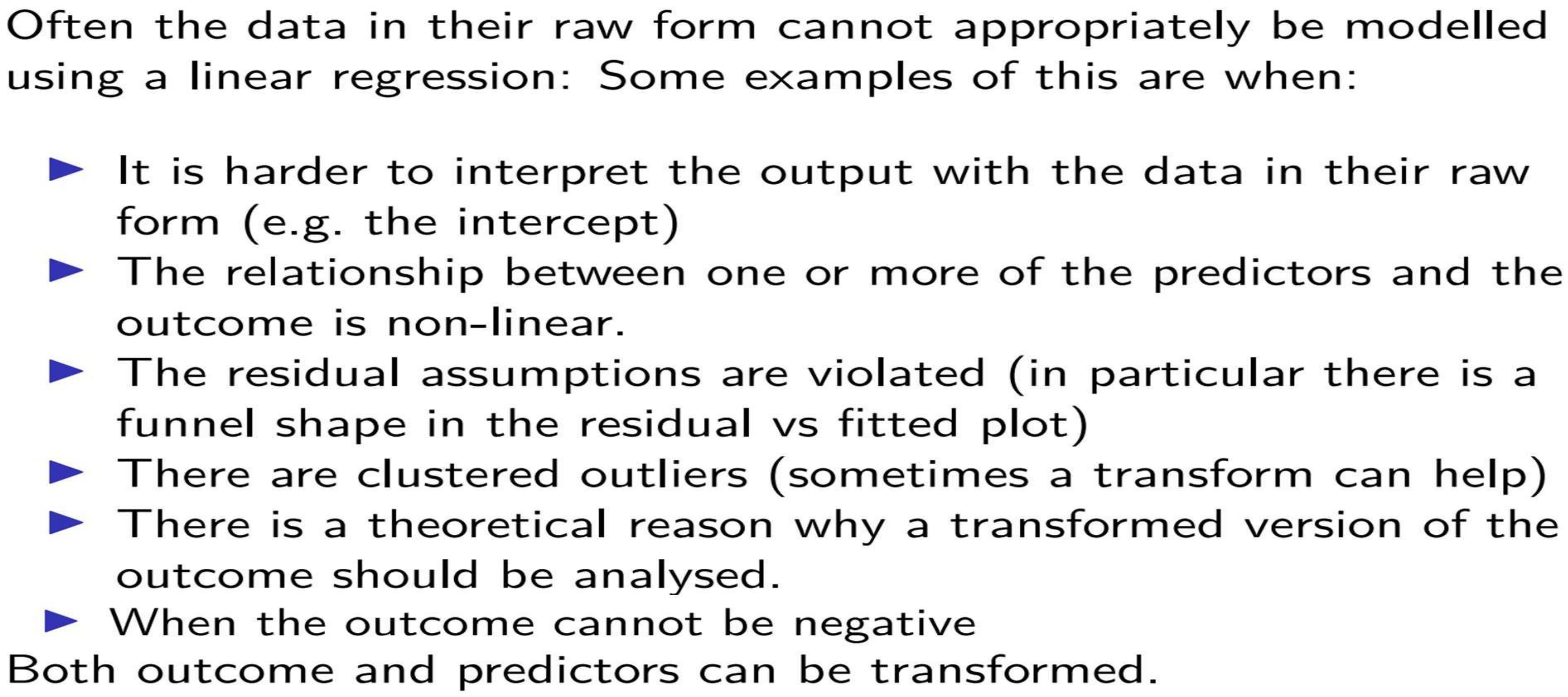
$$F = \frac{(\text{Regression SS})/p}{(\text{Residual SS})/(n-p-1)} \sim F_{p,\, n-p-1}$$

Adjusted R^2

- Adjusted R^2 is designed to be interpreted in the same way as the R^2 but for *multiple* rather than simple linear regression.
- It takes into account the loss of information incurred by adding a non-significant variable.
- There are two rules of thumb with using the Adjusted R^2 as a measure of model fit. For a good model:

1. It should be within 4% of the R^2 – otherwise there are probably too many non-significant predictors in the model
2. It should be close to 1 – same as the R^2.

Transformations



Often the data in their raw form cannot appropriately be modelled using a linear regression: Some examples of this are when:

- It is harder to interpret the output with the data in their raw form (e.g. the intercept)
- The relationship between one or more of the predictors and the outcome is non-linear.
- The residual assumptions are violated (in particular there is a funnel shape in the residual vs fitted plot)
- There are clustered outliers (sometimes a transform can help)
- There is a theoretical reason why a transformed version of the outcome should be analysed.
- When the outcome cannot be negative

Both outcome and predictors can be transformed.

Factors to consider

- Multicollinearity
- Variance inflation factor
- Overfitting

Thank you

Any Questions?