BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Digital Technologies and Value Creation

Dr. Philippe Blaettchen
Bayes Business School (formerly Cass)

www.bayes.city.ac.uk

Learning objectives of today

Goals: Recap the main tools in descriptive analytics

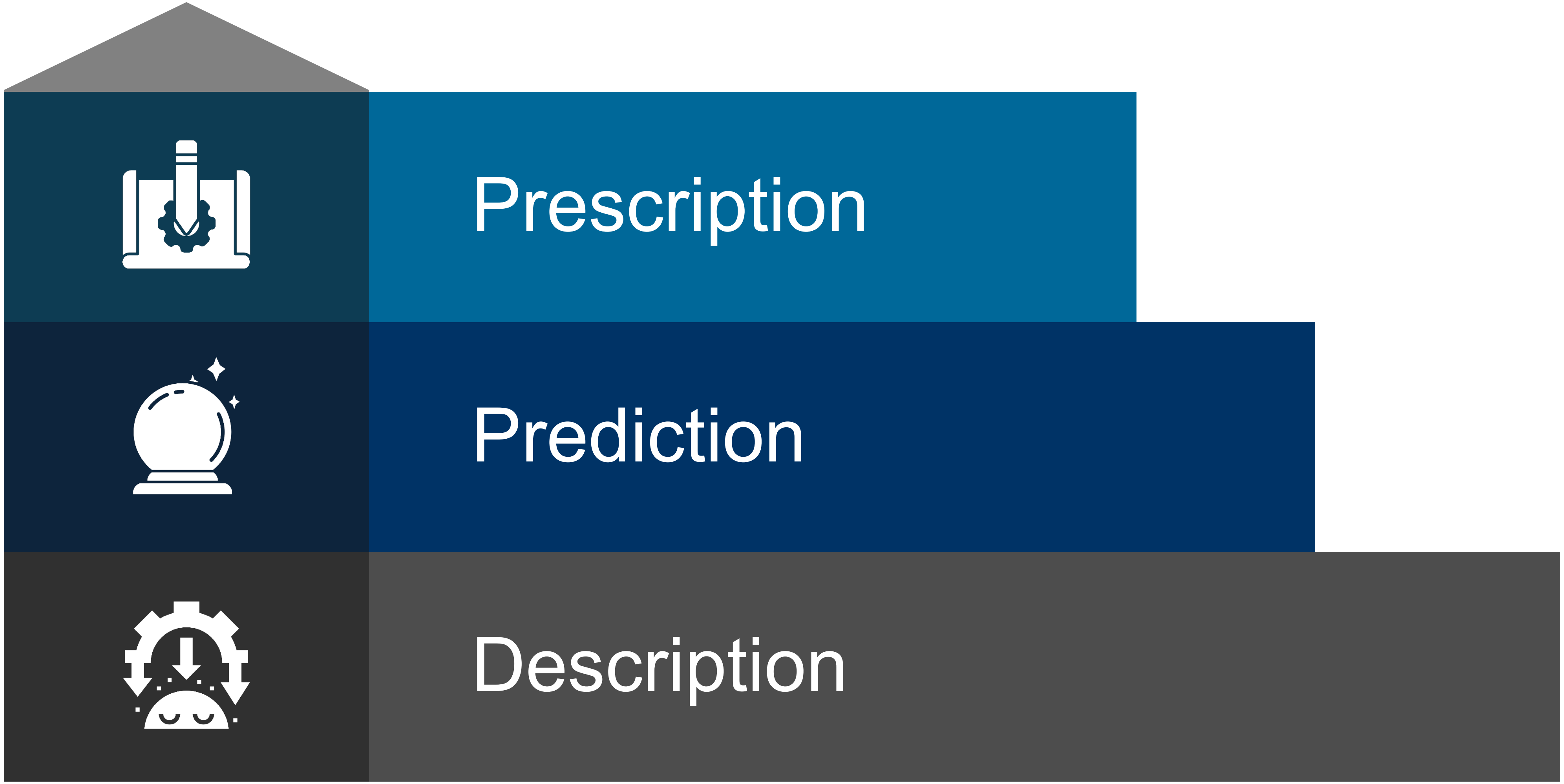
- Descriptive statistics
- Hypothesis testing

How will we do this?

- We consider a simple use case in people analytics, including a comprehensive (cleaned) dataset
- We will walk through some of the core descriptive tools theoretically
- We will then see how to implement these tools in Python

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Descriptive, predictive, or prescriptive – where is the value?



Descriptive (business) analytics

Using data to understand what's currently going on in or around the organization

1. Description (descriptive statistics)
2. Testing hunches, a.k.a. "Hypothesis testing" (bivariate and multi-variate associations)

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Our use-case

- It has about **18,000** employees
- The attrition rate of employees is quite high (**~14%** compared to industry average of **8-9%**).
  → How to fix it?
  - They have gathered data for calendar year 2020
  - Can you design a strategy to lower attrition?

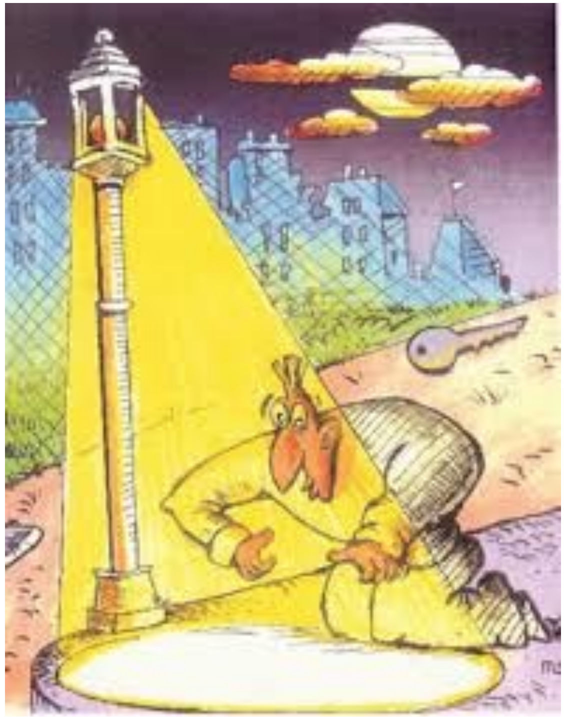
Descriptive statistics

1. Descriptive statistics

- Mean, Min, Max, Median, Std. Dev

- Helps understand
  - what the data is about
  - plausibility of data
  - missing data issues
  - how much variation there is in the data

- Remember to ask yourself: **what data are you NOT seeing in this sample?**



Let’s try it out

Hypothesis testing

2. Hypothesis testing

General idea: Does this association exist in the data?

- We have a hunch (hypothesis) about an explanation
- Null Hypothesis Significance Testing: Tells us the probability (across many samplings) of observing the association we see merely by chance, even if the true association in the population is zero.

2A. Bivariate association

- **t-tests** helps to understand association between a binary and a continuous/binary variable
  - E.g., is average salary (continuous variable) higher for those who exited vs. those who stayed (binary variable)?
  - Small p value says difference observed is unlikely to have arisen just by chance **if the true difference was zero**

- **Correlations** (-1,1) help to understand association between any two variables
  - E.g., is there a positive or negative correlation between age and salary? Between gender and exit? Between exit and age?
  - Small p value says correlation observed is unlikely to have arisen just by chance **if the true correlation was zero**.

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Let’s try it out

Correct interpretation of p (and confidence intervals)

The p-value is NOT

- the probability of unconditional replication
- the probability of null OR alternate hypotheses
- the probability of Type I error (i.e., probability of rejecting true null hypothesis)

Correct interpretation of p=x%:

Assuming there is no true effect, we would still obtain the observed difference or larger in x % of studies due to random sampling error.

"The probability of observing your data if the null is true."

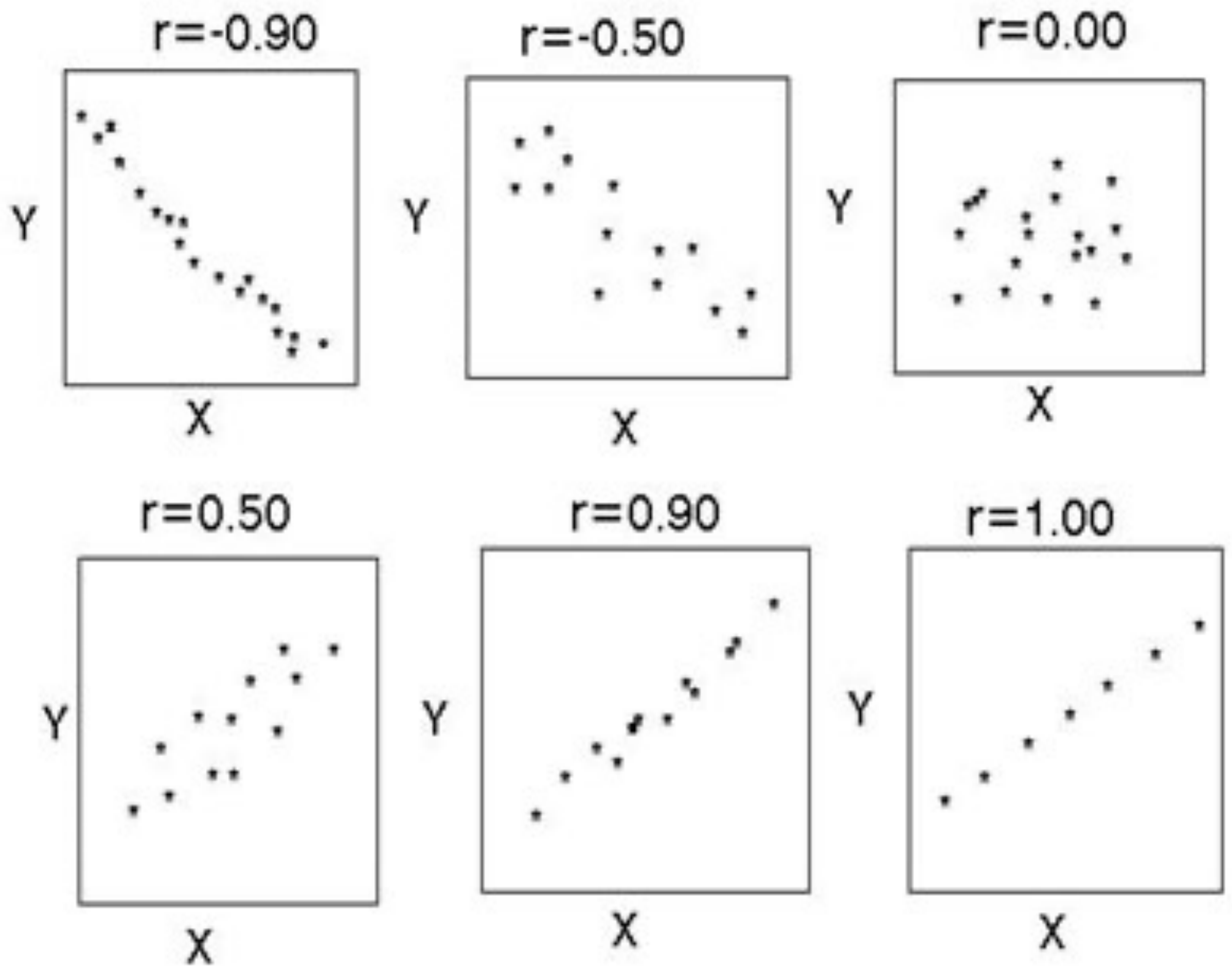
Correct interpretation of 95% confidence interval CI[a,b]:

If the true population parameter was equal to your sample parameter, in 95% of studies we conduct, the sample parameter would lie between a and b.

"The amount of uncertainty in your estimate"

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Correlation coefficients



BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

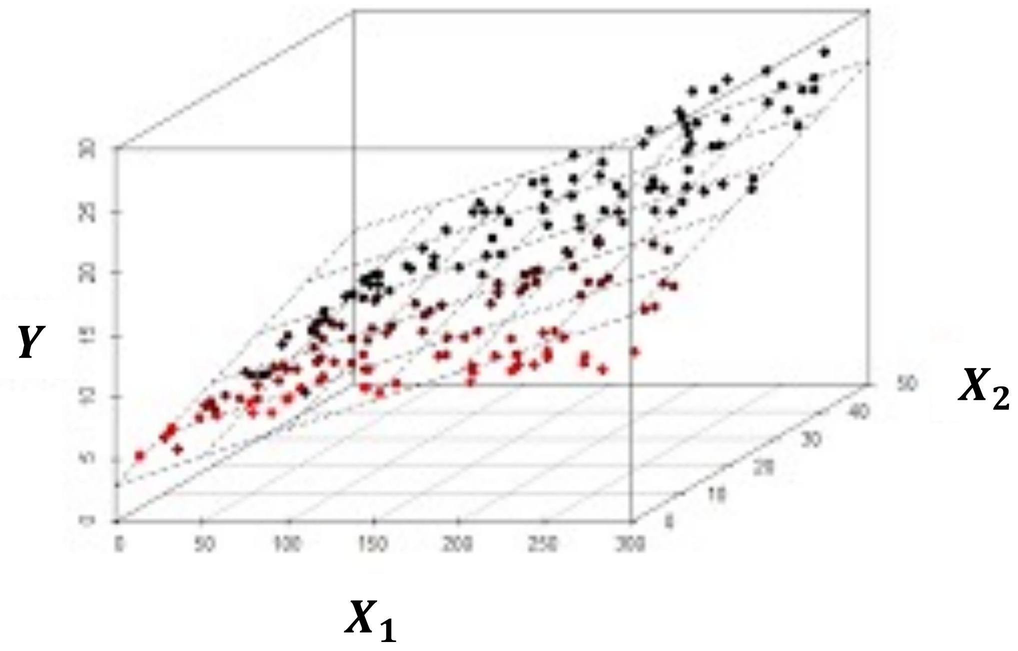
Multivariate association – OLS

2B. Multivariate association – OLS

- **Ordinary Least Squares (OLS)** helps to understand association between a dependent variable and several independent variables, **simultaneously**
  - E.g., is average salary higher for more educated employees, correcting for differences in age?
  - Small p value says coefficient observed is unlikely to have arisen just by chance **if the true difference was zero.**
  - A.k.a. "***regression model***" in data science jargon

BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

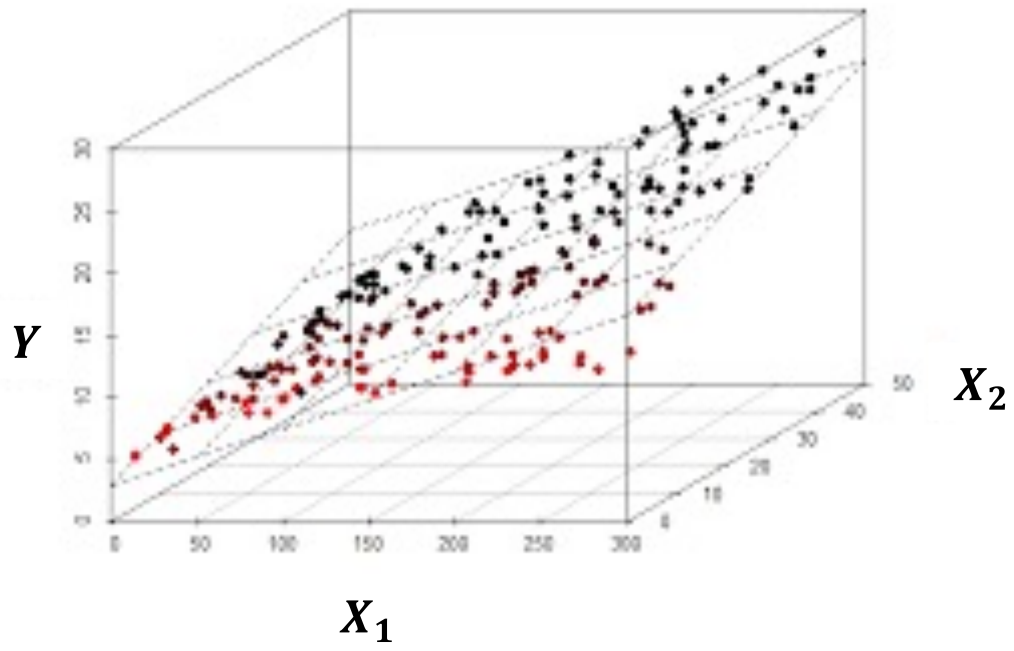
OLS "linear regression" model



Y = dependent variable
X_i = independent / explanatory variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

OLS "linear regression" model



How do we add this line/plane/hyperplane?

Input: datapoints (x_i, y_i)

Goal: find numbers $(\beta_0 = intercept, \beta_j = slope)$ such that sum of residuals squared $(y_1 - \beta_0 - \beta_1 \cdot x_{11} - \beta_2 \cdot x_{12} - \cdots)^2 + \cdots + (y_n - \beta_0 - \beta_1 \cdot x_{n1} - \beta_2 \cdot x_{n2} - \cdots)^2$ is as small as possible

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$

Linear regression in Python

- Two methods: statsmodels (stats viewpoint) and scikit (ML viewpoints)
- Statsmodels and scikit both have pros and cons: need to know how to modify examples given to you

```
X = df[['age','education']]
Y = df[['salary']]
X = sm.add_constant(X)
lm = sm.OLS(Y, X).fit()
print(lm.summary())
```

- Define your dependent variables in X, your independent variable in Y
  - With statsmodels, you have to add the constant manually
- Define your model, linear regression here
- Fit the linear regression
- Print properties of the model (R^2 and coefficients)

Linear regression in Python

- In general, how to read a summary of a regression? (via statsmodels here)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 salary   R-squared:                       0.119
Model:                            OLS   Adj. R-squared:                  0.118
Method:                 Least Squares   F-statistic:                     1219.
Date:                Tue, 12 Oct 2021   Prob (F-statistic):               0.00
Time:                        16:06:19   Log-Likelihood:                -59814.
No. Observations:               18132   AIC:                         1.196e+05
Df Residuals:                   18129   BIC:                         1.197e+05
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         19.9791      0.406     49.260      0.000      19.184      20.774
age            0.5058      0.010     49.339      0.000       0.486       0.526
education     -0.1110      0.074     -1.506      0.132      -0.255       0.033
==============================================================================
Omnibus:                     3594.862   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11932.879
Skew:                           0.998   Prob(JB):                         0.00
Kurtosis:                       6.437   Cond. No.                         319.
==============================================================================
```

Multiple R-squared: the closer to 1 the better the fit
Adjusted R-squared: takes into consideration number of vars

Coefficients are the coefficients of your line/plane:

$$salary = \beta_0 + \beta_1 \cdot age + \beta_2 \cdot education$$

Let’s try it out

python™

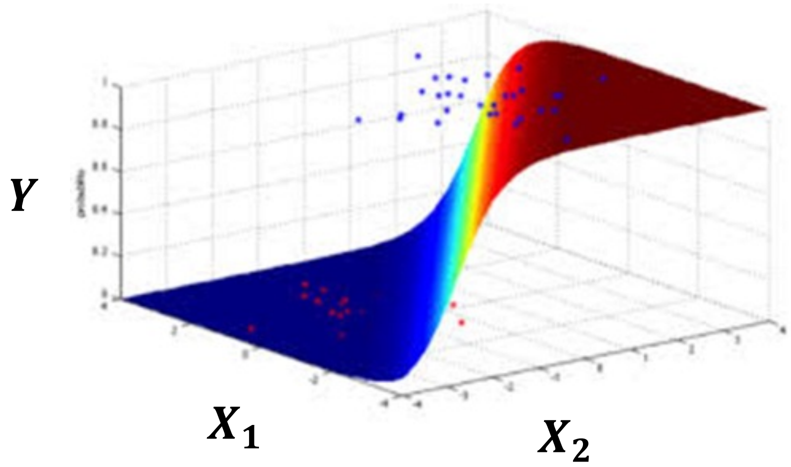
Multivariate association – logistic regression

2B. Multivariate association – logistic regression

- Sometimes, linear regression is **not good enough** – e.g., when the dependent variable is whether an employee leaves or not (binary)

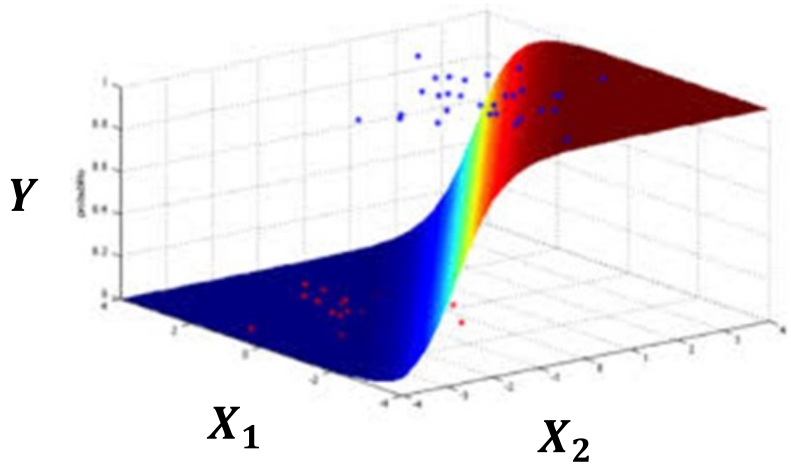
- **Logistic regressions** help to understand association between a binary dependent variable and several independent variables, **simultaneously**
  - E.g., Is exit more likely for older AND better paid employees ?
  - Small p value says correlation observed is unlikely to have arisen just by chance **if the true correlation was zero.**
  - A.k.a. "***classification model***" in data science jargon

Logistic “classification” model



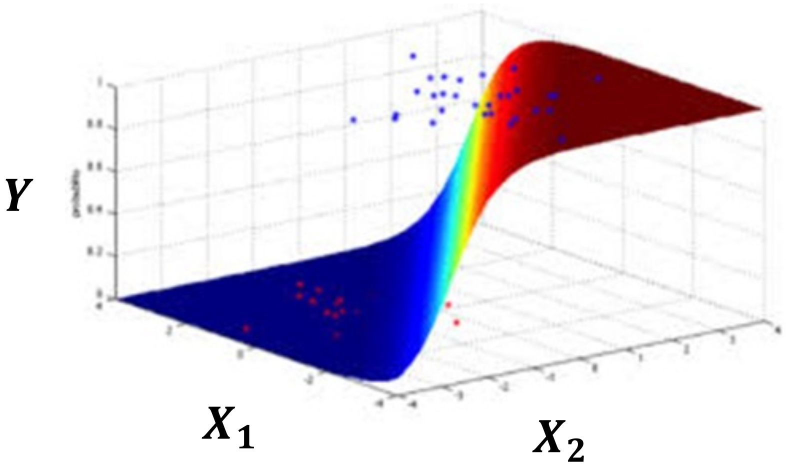
Input: datapoints (x_i, y_i)
Here; $y_i = 1$ (leaves) or 0 (doesn’t leave)

Logistic "classification" model



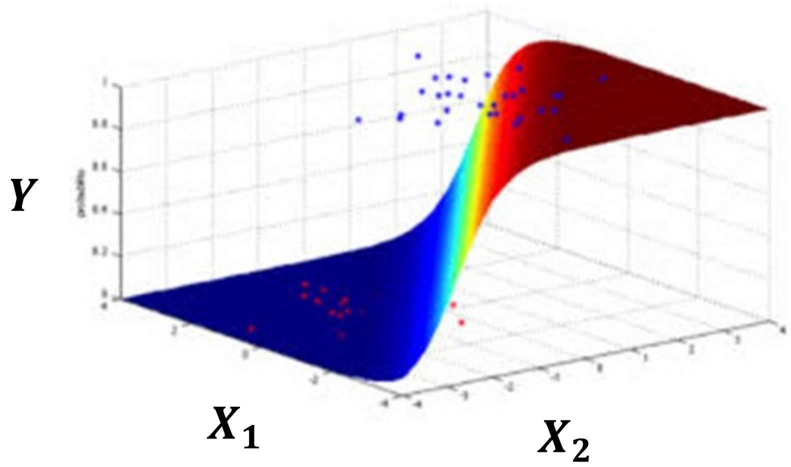
$$\underbrace{ln\left(\frac{P(Y)}{1-P(Y)}\right)}_{\text{The "log-odds" of belonging to a class}} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Logistic "classification" model



$$\frac{P(Y)}{1 - P(Y)} = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}$$

Logistic "classification" model



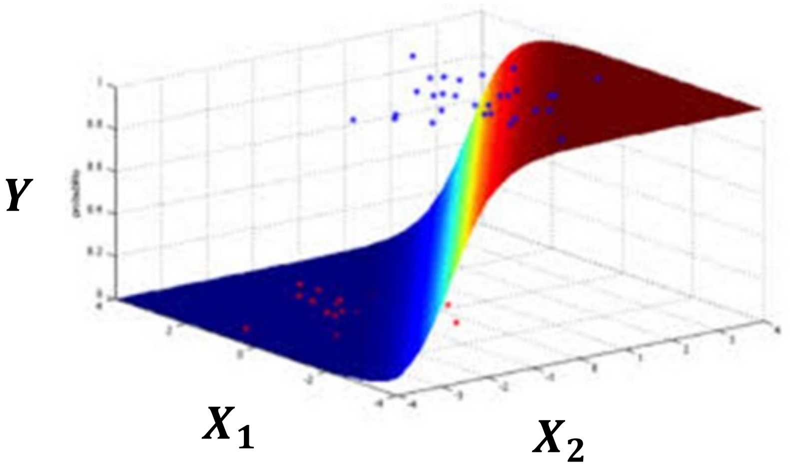
Advantage:

$$\frac{e^{\beta_0+\beta_1\cdot x_{i1}+\beta_2\cdot x_{i2}+\cdots}}{1+e^{\beta_0+\beta_1\cdot x_{i1}+\beta_2\cdot x_{i2}+\cdots}}$$

is a number between 0 and 1 (why?)

$$\boldsymbol{P(Y)=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon}}}$$

Logistic “classification” model



Goal: Find numbers (β_0, β_j) such that $\frac{e^{\beta_0+\beta_1 \cdot x_{i1}+\beta_2 \cdot x_{i2}+\cdots}}{1+e^{\beta_0+\beta_1 \cdot x_{i1}+\beta_2 \cdot x_{i2}+\cdots}}$ is as close as possible to y_i for all 1000 observations

$$P(Y) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\varepsilon}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\varepsilon}}$$

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Logistic regression in Python

- As before: statsmodels (stats viewpoint) and scikit (ML viewpoints)
- For now: statsmodels

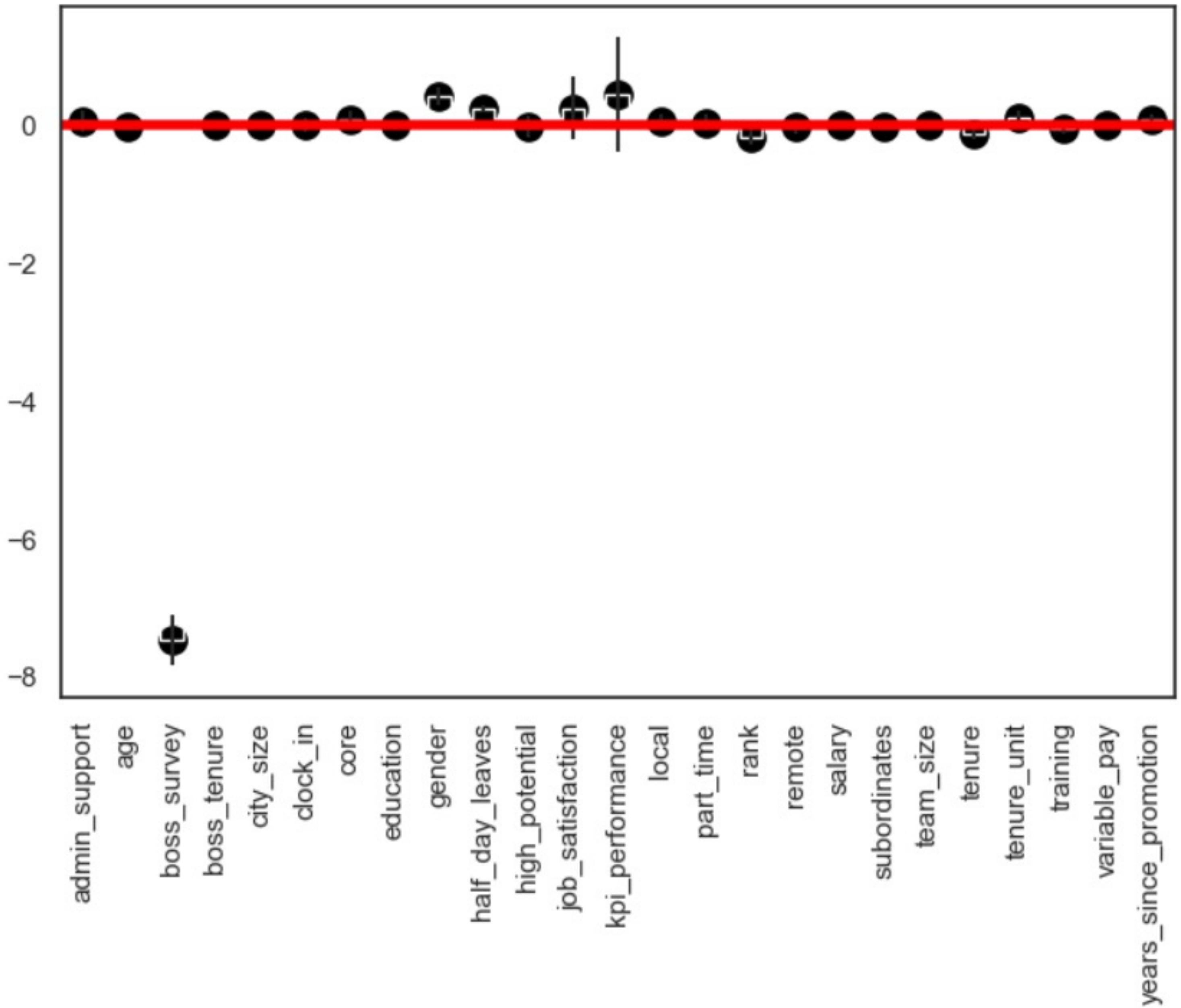
```
X = df.loc[:, df.columns != 'exit']
Y = df[["exit"]]
X = sm.add_constant(X)
lm = sm.OLS(Y,X).fit()
print (lm.summary())
```

The process is essentially the same as before:

- Define your dependent variables in X, your independent variable in Y
  - With statsmodels, you have to add the constant manually
- Define your model, linear regression here
- Fit the linear regression
- Print properties of the model (R^2 and coefficients)

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

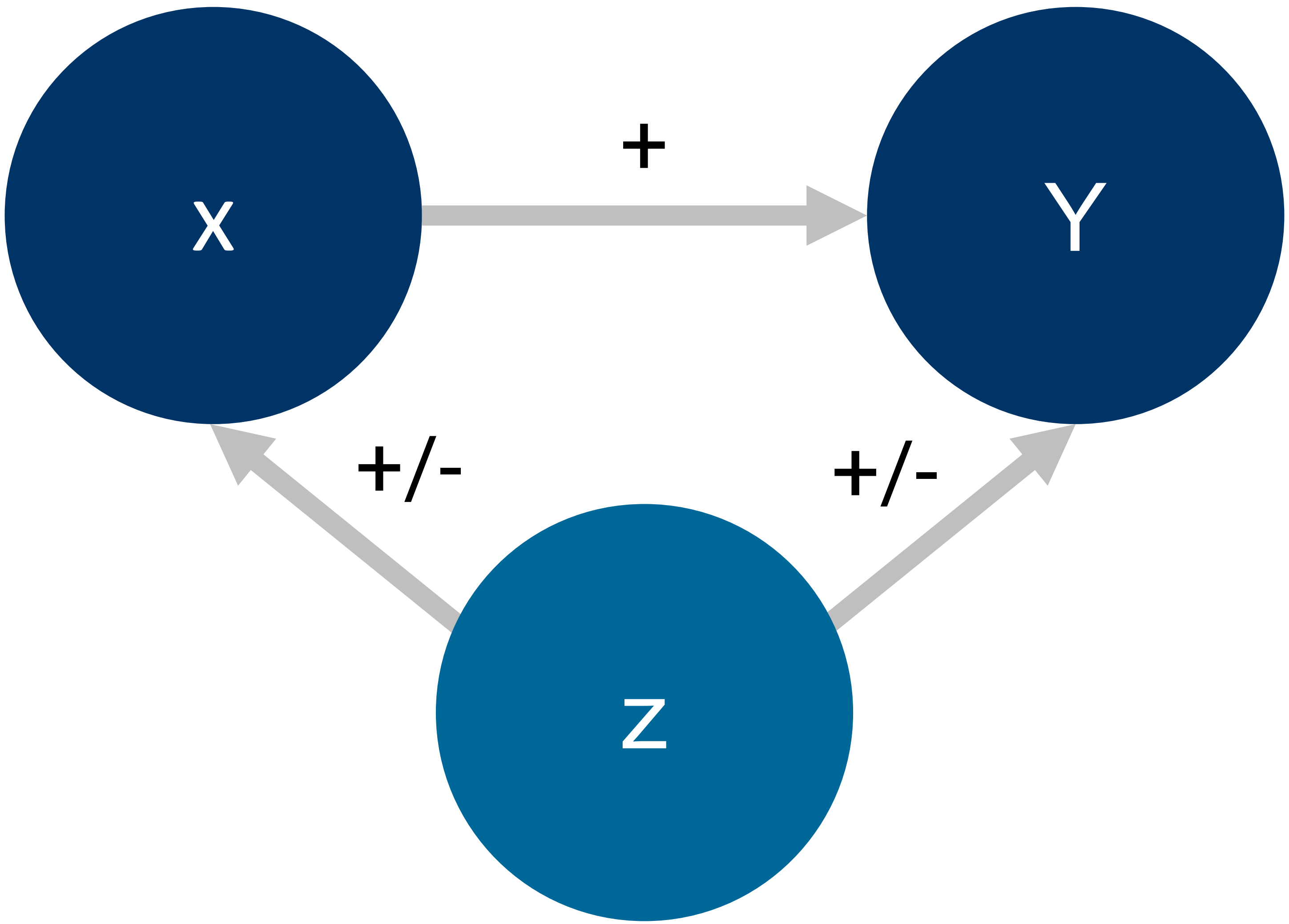
Let’s try it out



BAYES
BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Risks and limitations

Do storks deliver babies? The “hidden” variable effect



Warning about overfitting

- You have a hunch, test the association and get a low p-value
  - **Great!** Your hypothesis is supported! It is a good basis for guessing what might happen in the future.

- You hunt for associations with low p-values, treat these associations as results that can be useful to guess what might happen in the future
  - **Problem!** You might have found things that are only true in your sample, not in the population (remember, p values only indicate something about "average" samples)
  - A.k.a. "p-hacking" or "HARKING – hypothesizing after the results are known"

Summary

Summary

Statistical tools

- Descriptive statistics
- Hypothesis testing
  - Bi-variate associations
    - t-tests
    - correlations
  - Multi-variate associations
    - Regression (OLS)
    - Classification (Logistic)

Limitations

- Aggregate level, “on average” trends
- Correlation ≠ Causation
- Simple (linear) relationships
- Danger of misuse (overfitting)

BAYES BUSINESS SCHOOL
CITY, UNIVERSITY OF LONDON

Exercise

Helping Chimera

Answer the following questions, using the tools from descriptive analytics:

1. What is the economic impact of attrition on Chimera?
2. What do you think are some predictors of exit?
3. How would we test your hunches in the data? Try to implement your hypothesis tests, based on the video notebook

See you in class!