# Ordinary least squares regression I: The univariate linear regression.

Clement de Chaisemartin, UCSB

# Traders make predictions

- Traders, say oil traders, speculate on the price of oil.
- When they think the price of oil will go up, they buy oil.
- When they think the price will go down, they sell oil.
- To inform their buying / selling decisions, they need to predict whether the price will go up or down.
- To make their predictions, they can use the state of the economy today. E.g.: if world GDP is growing fast today, the price of oil should increase tomorrow.
- => traders need to use variables available to them to make predictions on a variable they do not observe: the price of oil tomorrow.

# Banks make predictions

- When someone applies for a loan, the bank needs to decide:
  - Whether they should give the loan to that person.
  - And if so, which interest rate they should charge that person.
- To answer these questions, the bank needs to predict the amount of the loan that this person will fail to reimburse. They will charge high interest rate to people who are predicted to fail to reimburse a large amount.
- To do so, they can use all the variables contained in application: gender, age, income, ZIP code…
- Can also use credit score of that person: FICO score, created by FICO company. All banks in US share information about their customers with FICO. Therefore, for each person FICO knows: total amount of debt, history of loans repayment… People with lots of debt and who often defaulted on their loan in the past get a low score, while people with little debt and no default get high score.
- Here as well, banks try to predict a variable they do not observe (amount of the loan the person will fail to reimburse) using variables that they observe (the variables in her application + FICO).

# Tech companies make predictions

- A reason why people prefer Gmail over other mailboxes is that Gmail is better than many mailboxes at sending directly spam emails into your trash box.
- They could ask a human to read the email and say whether it's a Spam or not. But that would be very costly and slow!
- Automated process: when a new email reaches your mailbox, Gmail needs to decide whether it should go into your trash because it's a Spam, or whether it should go into your regular mailbox.
- To do so, the computer can extract a number of variables from that email: number of words, email address of the sender, the specific words used in the email and how many times they occur…
- Based on these variables, it can try to predict whether the email is a real email or a spam.
- Here as well, Gmail tries to predict a variable they do not observe (whether that email is a Spam or not) using variables that they observe (number of words, email adress of the sender, the specific words used in the email…).

# Using variables we observe to make predictions on variables we do not observe.

- Many real world problems can be cast as using variables we observe to make predictions on variables we do not observe:
  - either because they will be realized in the future (e.g.: the amount that someone applying today for a one year to loan will fail to reimburse will only be known in one year from now)
  - or because observing them would be too costly (e.g.: assessing whether all the emails reaching all Gmail accounts everyday are spams or not).

We will study a variety of models one can use to make predictions.

- In all the following lectures, we are going to study how we can construct statistical models to make predictions.
- We will start by studying the simplest prediction model: the ordinary least squares (OLS) univariate linear regression.

# Roadmap

1. The OLS univariate linear regression function.
2. Estimating the OLS univariate linear regression function.
3. OLS univariate linear regression in practice.

# Set up and notation.

- We consider a population of $N$ units.
  - $N$ could be number of people who apply for a one-year loan with bank A during April 2018.
  - Or $N$ could be number of emails reaching all Gmail accounts in April 2018.
- Each unit $k$ has a variable $y_k$ attached to it that we do not observe. **We call this variable the dependent variable.**
  - In the loan example, $y_k$ is a variable equal to the amount of her loan applicant $k$ will fail to reimburse when her loan expires in April 2019.
  - In email example, $y_k$ is equal to 1 if email $k$ is a spam and 0 otherwise.
- Each unit $k$ also has 1 variable $x_k$ attached to it that we do observe. **We call this variable the independent variable.**
  - In the loan example, $x_k$ could be the FICO score of applicant $k$.
  - In the email example, $x_k$ could be a variable equal to 1 if the word "free" appears in the email.

# Are units with different values of $x_k$ likely to have the same value of $y_k$?

- Based on the value of $x_k$ of each unit, we want to predict her $y_k$.
- E.g.: in the loan example, we want to predict the amount that unit $k$ will fail to reimburse based on her FICO score.
- Assume that applicant 1 has a very high (good) credit score, while applicant 2 has a very low (bad) credit score.
- Do you think that applicant 1 and 2 will fail to reimburse the same amount on their loan?

# No!

- Based on the value of $x_k$ of each unit, we want to predict her $y_k$.
- E.g.: in the loan example, we want to predict the amount that unit $k$ will default on her loan based on her FICO score.
- Assume that applicant 1 has a very high (good) credit score, while applicant 2 has a very low (bad) credit score.
- Do you think that applicant 1 and 2 will fail to reimburse the same amount on their loan?
- No, applicant 2 is more likely to fail to reimburse a larger amount than applicant 1.
- Should you predict the same value of $y_k$ for applicants 1 and 2?

# No! Your prediction should be a function of $x_k$

- Based on the value of $x_k$ of each unit, we want to predict her $y_k$.
- E.g.: in the loan example, we want to predict the amount that unit $k$ will default on her loan based on her FICO score.
- Assume that applicant 1 has a very high (good) credit score, while applicant 2 has a very low (bad) credit score.
- Should you predict the same value of $y_k$ for applicants 1 and 2?
- No! If you want your prediction to be accurate, you should predict a higher value of $y_k$ for applicant 2 than for applicant 1.
- Your prediction should a function of $x_k$, $f(x_k)$.
- **In these lectures, we focus on predictions which are a linear function of $x_k$: $f(x_k) = ax_k$, for some real number $a$.**
- Which measure can you use to assess whether $ax_k$ is a good prediction of $y_k$? Discuss this question with your neighbor for 1 minute.

# iClicker time

- To assess whether $ax_k$ is a good prediction of $y_k$, we should use:

$a)\ y_k - ax_k$

$b)\ ax_k + y_k$

# $y_k - ax_k$!

- Based on the value of $x_k$ of each unit, we want to predict her $y_k$.
- Our prediction should a function of $x_k$, $f(x_k)$. We focus on predictions which are a linear function of $x_k$: $f(x_k) = ax_k$, for some real number $a$.
- Which measure can you use to assess whether $ax_k$ is a good prediction?
- $y_k - ax_k$ , the difference between your prediction and $y_k$.
- In the loan example, if $y_k - ax_k$ is large and positive, our prediction is much below the amount applicant $k$ will fail to reimburse.
- If $y_k - ax_k$ is large and negative, our prediction is much above the amount person $k$ will fail to reimburse.
- Large positive or negative values of $y_k - ax_k$ mean bad prediction.
- $y_k - ax_k$ close to 0 means good prediction.

# iClicker time

- Which of the following 3 possible values of $a$ should we choose to ensure that $ax_k$ predicts well $y_k$ in the population?

a) The value of $a$ that maximizes $\sum_{k=1}^{N}(y_k - ax_k)$

b) The value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)$ .

c) The value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$.

Minimizing $\sum_{k=1}^{N}(y_k - ax_k)$ won't work!

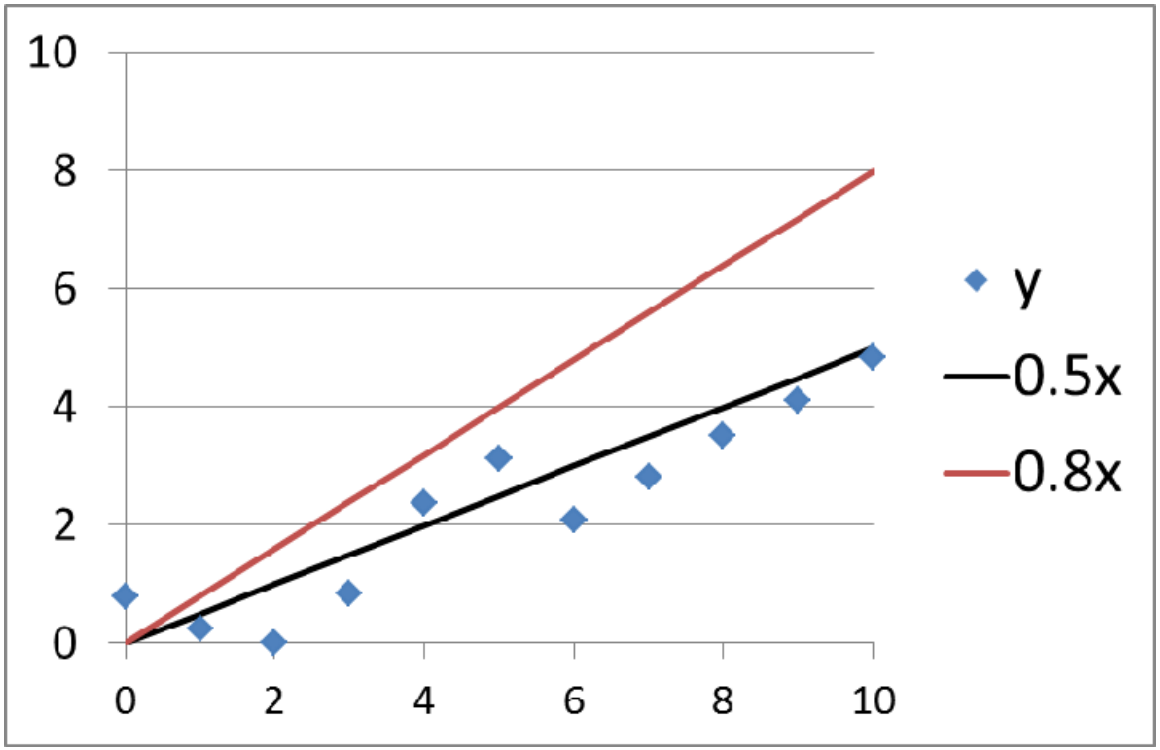
- Minimizing $\sum_{k=1}^{N}(y_k - ax_k)$ means we try to avoid positive prediction errors, but we also try to make the largest possible negative prediction errors!
- Not a good idea: we will systematically overestimate $y_k$.
- We want a criterion that deals symmetrically with positive and negative errors: we want to avoid both positive and negative errors.

Answer: find the value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$

- $\sum_{k=1}^{N}(y_k - ax_k)^2$ is positive. => minimizing it = same thing as making it as close to 0 as possible.
- If $\sum_{k=1}^{N}(y_k - ax_k)^2$ is as close to 0 as possible, means that the sum of the squared value of our prediction errors is as small as possible.
- => we make small errors. That's good, that's what we want!

# Which prediction function is the best?

- Population has 11 units. $x_k$ and $y_k$ of those 11 units are shown on the graph: blue dots.
- Two linear prediction functions for $y_k$: $0.5x_k$ and $0.8x_k$.
- Which one is the best? Discuss this 1mn with your neighbour.



# iClicker time

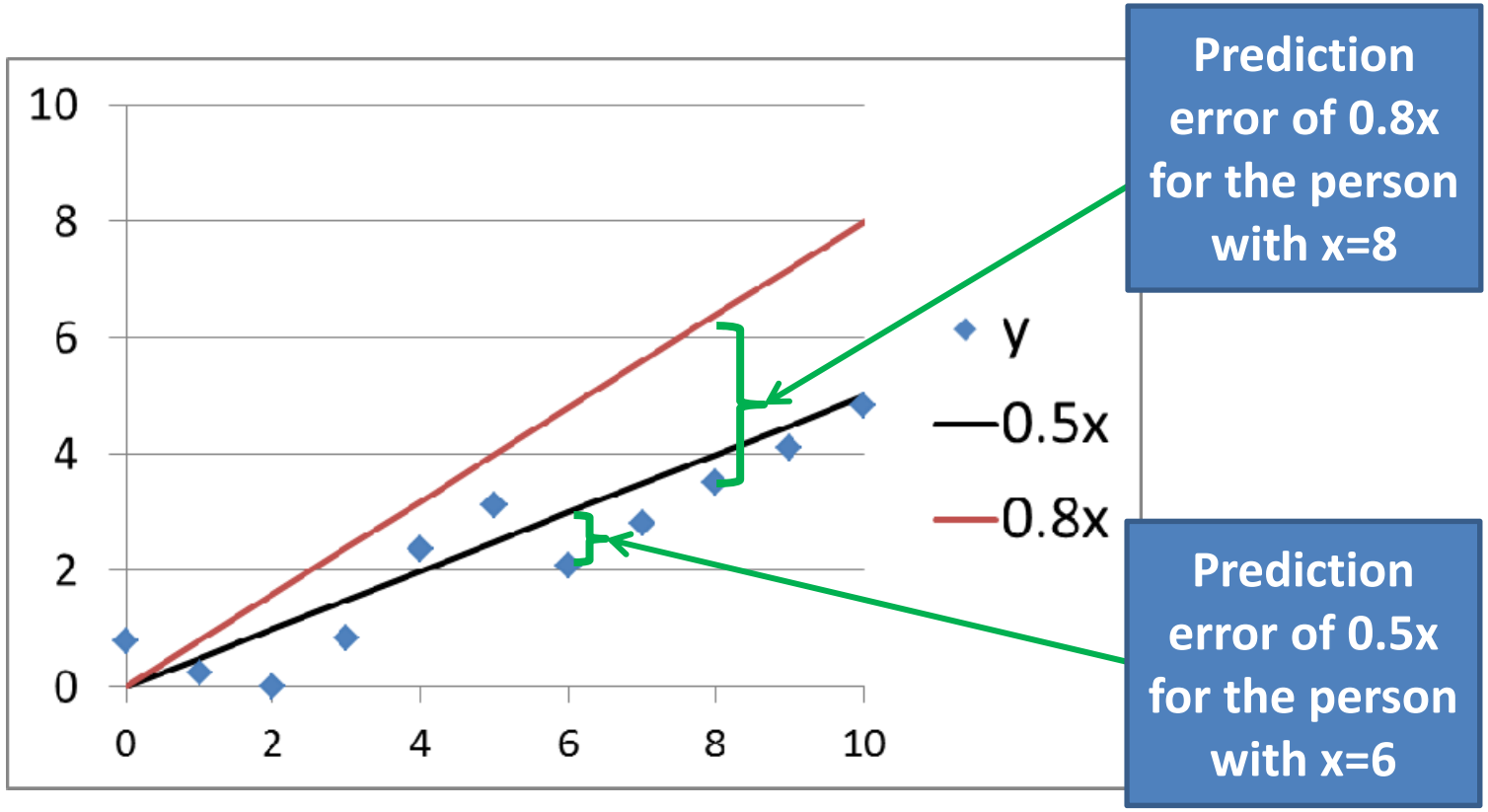
- On the previous slide, which function of $x_k$ gives the best prediction for $y_k$:

a) $0.5x_k$

b) $0.8x_k$

# $0.5x_k$ is the best prediction function!

- It is the function for which the sum of the squared of the prediction errors are the smallest.



# The OLS univariate linear regression function in the population.

- Let

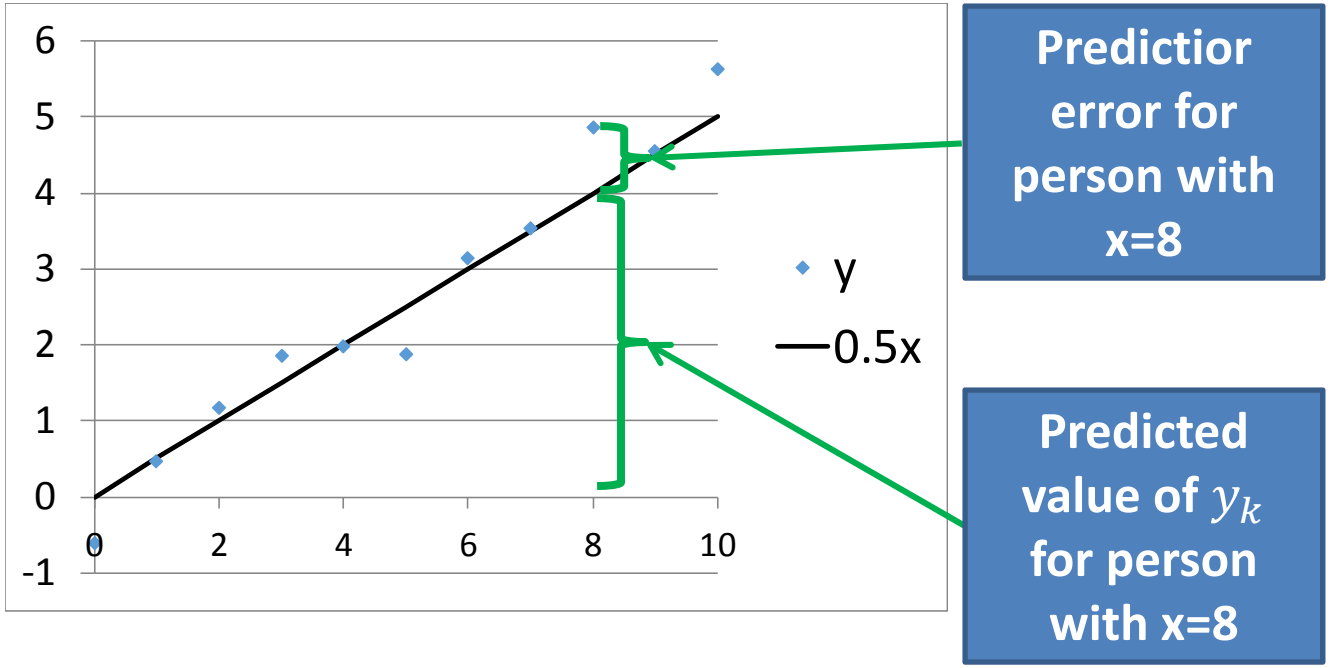
$$\alpha = argmin_{a \in R} \sum_{k=1}^{N} (y_k - a x_k)^2$$

- We call $\alpha x_k$ the **ordinary least squares (OLS) univariate linear regression function of $y_k$ on $x_k$ in the population.**
- Least squares: because $\alpha x_k$ minimizes the sum of the squared difference between $y_k$ and $a x_k$.
- Ordinary: because there are fancier way of doing least squares.
- Univariate: because there is only one independent variable in the regression, $x_k$.
- Linear: because the regression function is a linear function of $x_k$.
- In the population: because we use the $y_k$s and $x_k$s of all the $N$ units in the population.
- Shortcut: **OLS regression of $y_k$ on $x_k$ in the population**.

# Decomposing $y_k$ between predicted value and residual.

- $\alpha$: coefficient of $x_k$ in the OLS regression of $y_k$ on $x_k$ in the population.
- Let $\tilde{y}_k = \alpha x_k$. $\tilde{y}_k$ is the predicted value for $y_k$ according to the OLS regression of $y_k$ on $x_k$ in the population.
- Let $e_k = y_k - \tilde{y}_k$. $e_k$: error we make when we use OLS regression in the population to predict $y_k$.
- We have $y_k = \tilde{y}_k + e_k$.

$y_k$ =predicted value + error.



# Finding a formula for $\alpha$ when $N = 2$.

- Assume for a minute that $N = 2$: there are only two units in the population.
- Then $\alpha$ is the value of $a$ that minimizes $(y_1 - ax_1)^2 + (y_2 - ax_2)^2$.
- Find a formula for $\alpha$, as a function of $y_1$, $x_1$, $y_2$, and $x_2$. You have 3 minutes to try to find the answer. Hint: you need to compute the derivative of $(y_1 - ax_1)^2 + (y_2 - ax_2)^2$ with respect to $a$, and then $\alpha$ is the value of $a$ for which that derivative is equal to 0.

# iClicker time

- If $N = 2$, $\alpha$ is equal to:

a) $\frac{x_1 y_1 + x_2 y_2}{x_1 + x_2}$

b) $\frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$

c) $\frac{{x_1}^2 y_1 + {x_2}^2 y_2}{x_1 + x_2}$

# When $N = 2$, $\alpha = \frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$.

- If $N = 2$, $\alpha$ is the value of $a$ that minimizes $(y_1 - ax_1)^2 + (y_2 - ax_2)^2$.
- The derivative of that function wrt to $a$ is:

$-x_1 2(y_1 - ax_1) - x_2 2(y_2 - ax_2)$.

- Let's find value of $a$ for which derivative = 0.

$$-2x_1(y_1 - ax_1) - 2x_2(y_2 - ax_2) = 0$$
$$\text{i}if\ -2x_1 y_1 + 2a{x_1}^2 - 2x_2 y_2 + 2a{x_2}^2 = 0$$
$$\text{i}if\ 2a({x_1}^2 + {x_2}^2) = 2(x_1 y_1 + x_2 y_2)$$
$$\text{i}if\ a = \frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}.$$

- Second line of derivation shows that derivative increasing in $a$. => if $a < \frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$, derivative negative. If $a > \frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$, derivative positive.
- **Function reaches minimum at $\frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$. => $\alpha = \frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$.**

# Reminder: P4Sum

- P4Sum: let $f_1(a)$, $f_2(a)$,…, $f_N(a)$ be $N$ functions of $a$ which are all differentiable wrt to $a$. Let $f_1{}'(a)$,$f_2{}'(a)$,…,$f_N{}'(a)$ denote their derivatives. Then, $\sum_{k=1}^{N} f_k(a)$ is differentiable wrt to $a$, and its derivative is $\sum_{k=1}^{N} f_k{}'(a)$.
- In words: the derivative of a sum is the sum of its derivatives.

# Finding a formula for $\alpha$ for any value of $N$.

- Let's get back to the general case where $N$ is left unspecified.
- Remember, $\alpha$ is the value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$.
- Find a formula for $\alpha$, as a function of $y_1,\ldots, y_N$ and $x_1,\ldots,x_N$. You have 3 minutes to find the answer. Hint: you need to compute the derivative of $\sum_{k=1}^{N}(y_k - ax_k)^2$with respect to $a$, and then $\alpha$ is the value of $a$ for which that derivative is equal to 0.

# iClicker time

- $\alpha$ is equal to:

a) $\frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} x_k}$

b) $\frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$

c) $\frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$

$$\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}.$$

- $\alpha$ minimizes $\sum_{k=1}^{N}(y_k - a x_k)^2$. Derivative wrt to $a$ is: $\sum_{k=1}^{N}[-x_k 2(y_k - a x_k)]$. Why?
- Let's find value of $a$ for which derivative = 0.

$$\sum_{k=1}^{N}[-2x_k(y_k - a x_k)] = 0$$

$$\text{i}if \sum_{k=1}^{N}[-2x_k y_k + 2a{x_k}^2] = 0$$

$$\text{i}if \sum_{k=1}^{N} -2x_k y_k + \sum_{k=1}^{N} 2a{x_k}^2 = 0$$

$$\text{i}if -2\sum_{k=1}^{N} x_k y_k + 2a\sum_{k=1}^{N} {x_k}^2 = 0$$

$$\text{i}if 2a\sum_{k=1}^{N} {x_k}^2 = 2\sum_{k=1}^{N} x_k y_k$$

$$\text{i}if\ a = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}.$$

- **Function reaches minimum at $\frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$. => $\boldsymbol{\alpha} = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$.**

# What you need to remember

- Population of $N$ units. Each unit $k$ has 2 variables attached to it: $y_k$ is a variable we do not observe, $x_k$ is a variable we observe.
- We want to predict the $y_k$ of each unit based on her $x_k$.
- E.g.: a bank wants to predict the amount an applicant will fail to reimburse on her loan based on her FICO score.
- Our prediction should be function of $x_k$, $f(x_k)$.
- For now, focus on linear functions of $x_k$: $ax_k$ for some number $a$.
- Good prediction should be such that $y_k - ax_k$, difference between prediction and $y_k$, is as small as possible for most units.
- The best value of $a$ is the one that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$.
- We call that value $\alpha$, and we call $\alpha x_k$ the OLS univariate linear regression function of $y_k$ on $x_k$.
- If $N = 2$ $\alpha = \frac{x_1 y_1 + x_2 y_2}{{x_1}^2 + {x_2}^2}$. You should know how to prove that.
- In general, $\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$. You should know how to prove that.

# Roadmap

1. The OLS univariate linear regression function.
2. Estimating the OLS univariate linear regression function.
3. OLS univariate linear regression in practice.

# Can we compute $\alpha$?

- Our prediction for $y_k$ based on a univariate linear regression is $\alpha x_k$, the univariate linear regression function.
- => to be able to make a prediction for a unit's $y_k$ based on her $x_k$, we need to know the value of $\alpha$.
- Under the assumptions we have made so far, can we compute $\alpha$? Discuss this question with your neighbor during 1 minute.

# iClicker time

- Under the assumptions we have made so far, can we compute $\alpha$?

a) Yes

b) No

# We do not observe the $y_k$s, => we cannot compute $\alpha$

- Remember, we have assumed that we observe the $x_k$s of everybody in the population (e.g. applicants' FICO scores) but not the $y_k$s (e.g. the amount that a person applying for a one-year loan in April 2018 will fail to reimburse in April 2019 when that loan expires).
- => we cannot compute $\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$.

## But we can estimate $\alpha$ if we observe the $y_k$s of a sample of the population.

- But we can estimate $\alpha$ if we observe the $y_k$s of a sample of the population.
- E.g.: in the Gmail example, we could select a random sample of emails, and ask a human to determine whether those emails are spams or not.

# Randomly sampling one unit.

- Assume we randomly select one unit in the population, and we measure the dependent and the independent variable of that unit.
- E.g.: we randomly select one email out of all the emails reaching Gmail accounts on May, 1st, 2018, and we look whether this is a spam or not, and whether it contains the word "free" or not.
- Let $Y_1$ and $X_1$ respectively denote the value of the dependent and of the independent variable for that randomly selected unit.
- $Y_1$ and $X_1$ are random variables, because their values depend on which unit of the population we randomly select.
- If we select the $34$th unit in the population, $Y_1 = y_{34}$ and $X_1 = x_{34}$.
- Each unit in the population has the same probability, $\frac{1}{N}$, of being selected.
- What is the value of $E(X_1 Y_1)$? Hint: $E(X_1 Y_1)$ is a function of all the $y_k$s and of all the $x_k$s. Discuss this question with your neighbor during 2mns.

# iClicker time

- Assume we randomly select one unit in the population, and we measure the dependent and the independent variable of that unit.
- Let $Y_1$ and $X_1$ respectively denote the value of the dependent and of the independent variable for that randomly selected unit.
- $Y_1$ and $X_1$ are random variables, because their values depend on which unit of the population of the population we randomly select.
- Each unit in the population has a probability $\frac{1}{N}$ of being selected.
- What is the value of $E(X_1 Y_1)$?

*a)* $E(X_1 Y_1) = x_k y_k$

*b)* $E(X_1 Y_1) = \frac{1}{N} \sum_{k=1}^{N} x_k y_k$

*c)* $E(X_1 Y_1) = \sum_{k=1}^{N} x_k y_k$

$$E(X_1 Y_1) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$$

- $X_1 Y_1$ is equal to:
  - $x_1 y_1$ if the first individual in the population is selected, which has a probability $\frac{1}{N}$ of happening
  - $x_2 y_2$ if the second individual in the population is selected, which has a probability $\frac{1}{N}$ of happening
  - …
  - $x_N y_N$ if the $N$th individual in the population is selected, which has a probability $\frac{1}{N}$ of happening
- Therefore, $E(X_1 Y_1) = \sum_{k=1}^{N} \frac{1}{N} x_k y_k = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$.
- What is the value of $E\left({X_1}^2\right)$? Discuss this question with your neighbor during 1mn.

# iClicker time

- We randomly select one unit, and we measure the dependent and the independent variable of that unit.
- Let $Y_1$ and $X_1$ respectively denote the value of the dependent and of the independent variable for that randomly selected unit.
- $Y_1$ and $X_1$ are random variables, because their values depend on which unit of the population of the population we randomly select.
- Each unit in the population has a probability $\frac{1}{N}$ of being selected.
- What is the value of $E({X_1}^2)$?

*a)* $E({X_1}^2) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$

*b)* $E({X_1}^2) = \frac{1}{N}\sum_{k=1}^{N} x_k$

$$E\left({X_1}^2\right) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$$

- ${X_1}^2$ is equal to:
  - ${x_1}^2$ if the first individual in the population is selected, which has a probability $\frac{1}{N}$ of happening
  - ${x_2}^2$ if the second individual in the population is selected, which has a probability $\frac{1}{N}$ of happening
  - …
  - ${x_N}^2$ if the $N$th individual in the population is selected, which has a probability $\frac{1}{N}$ of happening
- Therefore, $E\left({X_1}^2\right) = \sum_{k=1}^{N} \frac{1}{N} {x_k}^2 = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$ .

# Randomly sampling $n$ units.

- We randomly draw $n$ units with replacement from the population, and we measure the dependent and the independent variable of those $n$ units.
- For every $i$ included between 1 and $n$, $Y_i$ and $X_i$ = value of the dependent and of the independent variable of the $i$th unit we randomly select.
- The $Y_i$s and $X_i$s are independent and identically distributed.
- For every $i$ included between 1 and $n$, $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left({X_i}^2\right) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$ .

# A method to estimate $\alpha$.

- We want to use the $Y_i$s and the $X_i$s to estimate $\alpha$.
- Remember: $\alpha$ is the value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$.
- => to estimate $\alpha$, we could use $\hat{\alpha}$, the value of $a$ that minimizes $\sum_{i=1}^{n}(Y_i - aX_i)^2$.
- Instead of finding the value of $a$ that minimizes the sum of squared prediction errors in the population, find value of $a$ that minimizes the sum of squared prediction errors in the sample.
- Intuition: if we find a method to predict well the dependent variable in the sample, that method should also work well in the full population, as our sample is representative of the population.

# The OLS regression function in the sample.

- Let

$$\hat{\alpha} = argmin_{a \in R} \sum_{i=1}^{n} (Y_i - aX_i)^2$$

- We call $\hat{\alpha}X_i$ the OLS regression function of $Y_i$ on $X_i$ **in the sample.**
- In the sample: because we only use the $Y_i$s and $X_i$s of the $n$ units in the sample we randomly draw from the population.
- $\hat{\alpha}$: coefficient of $X_i$ in the OLS regression of $Y_i$ on $X_i$ **in the sample**.
- Let $\hat{Y}_i = \hat{\alpha}X_i$. $\hat{Y}_i$ is the predicted value for $Y_i$ according to the OLS regression of $Y_i$ on $X_i$ **in the sample**.
- Let $\hat{e}_i = Y_i - \hat{Y}_i$. $\hat{e}_i$: error we make when we use OLS regression **in the sample** to predict $\hat{Y}_i$.
- We have $Y_i = \hat{Y}_i + \hat{e}_i$.
- Find a formula for $\hat{\alpha}$, the value of $a$ that minimizes $\sum_{i=1}^{n}(Y_i - aX_i)^2$. Hint: differentiate this function wrt to $a$ and find the value of $a$ that cancels the derivative.

# iClicker time

- The value of $a$ that minimizes $\sum_{i=1}^{n}(Y_i - aX_i)^2$ is:

*a)* $\frac{\sum_{i=1}^{n} {X_i}^2 Y_i}{\sum_{i=1}^{n} X_i}$

*b)* $\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}$

*c)* $\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i}$

## Value of $a$ minimizing $\sum_{i=1}^{n}(Y_i - aX_i)^2$ is $\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}$

- Derivative wrt to $a$ of $\sum_{i=1}^{n}(Y_i - aX_i)^2$ is: $\sum_{i=1}^{n}[-X_i 2(Y_i - aX_i)]$. Why?
- Let's find value of $a$ for which derivative = 0.

$$\sum_{i=1}^{n}[-X_i 2(Y_i - aX_i)] = 0$$

$$if \sum_{i=1}^{n}\left[-2X_i Y_i + 2a{X_i}^2\right] = 0$$

$$if \sum_{i=1}^{n} -2X_i Y_i + \sum_{i=1}^{n} 2a{X_i}^2 = 0$$

$$if -2\sum_{i=1}^{n} X_i Y_i + 2a\sum_{i=1}^{n} {X_i}^2 = 0$$

$$if 2a\sum_{i=1}^{n} {X_i}^2 = 2\sum_{i=1}^{n} X_i Y_i$$

$$if\ a = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}.$$

$$\boldsymbol{\hat{\alpha} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}}$$

# Reminder: the law of large numbers.

- LLN: Let $Z_1$,…, $Z_n$ be $n$ iid random variables, and let $m$ denote their expectation.

  $$\lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = m.$$

- When the sample size grows, the average of $n$ iid random variables converges towards their expectation.

# $\hat{\alpha}$ converges towards $\alpha$ when the sample size grows.

- We randomly draw $n$ units with replacement from the population, and we measure the dependent and the independent variable of those $n$ units.
- For every $i$ included between 1 and $n$, $Y_i$ and $X_i$ = value of the dependent and of the independent variable of the $i$th unit we randomly select.
- Because the $n$ units are drawn with replacement, $(Y_i, X_i)$ are iid, and therefore the $X_i Y_i$s are iid, and the ${X_i}^2$ s are also iid.
- For every $i$ included between 1 and $n$, $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left({X_i}^2\right) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$ .
- $\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$ and $\hat{\alpha} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}$.
- Use the law of large numbers to show that $\lim_{n \to +\infty} \hat{\alpha} = \alpha$. Hint: you need to use the fact that $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left({X_i}^2\right) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$ .

# iClicker time

- Which of following two arguments is correct:

a) The law of large numbers implies that $\lim_{n\to+\infty} \sum_{i=1}^{n} X_i Y_i = E(X_i Y_i)$ and $\lim_{n\to+\infty} \sum_{i=1}^{n} {X_i}^2 = E\left({X_i}^2\right)$. We have $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left({X_i}^2\right) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$. Therefore, $\lim_{n\to+\infty} \hat{\alpha} = \alpha$.

b) The law of large numbers implies that $\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i = E(X_i Y_i)$ and $\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} {X_i}^2 = E\left({X_i}^2\right)$. We have $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left({X_i}^2\right) = \frac{1}{N}\sum_{k=1}^{N} {x_k}^2$. Therefore, $\lim_{n\to+\infty} \hat{\alpha} = \alpha$.

# The second argument is correct

- We randomly draw $n$ units with replacement from the population, and we measure the dependent and the independent variable of those $n$ units.
- For every $i$ included between 1 and $n$, $Y_i$ and $X_i$ = value of the dependent and of the independent variable of the $i$th unit we randomly select.
- Because the $n$ units are drawn with replacement, $(Y_i, X_i)$ are iid.
- For every $i$ included between 1 and $n$, $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left(X_i{}^2\right) = \frac{1}{N}\sum_{k=1}^{N} x_k{}^2$ .
- The law of large numbers implies that $\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i = E(X_i Y_i)$ and $\lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^{n} X_i{}^2 = E\left(X_i{}^2\right)$.
- Moreover, We have $E(X_i Y_i) = \frac{1}{N}\sum_{k=1}^{N} x_k y_k$ and $E\left(X_i{}^2\right) = \frac{1}{N}\sum_{k=1}^{N} x_k{}^2$ .
- Therefore:

$$\lim_{n\to+\infty} \hat{\alpha} = \lim_{n\to+\infty} \frac{\sum_{i=1}^{n} \boldsymbol{X_i Y_i}}{\sum_{i=1}^{n} \boldsymbol{X_i}^2} = \lim_{n\to+\infty} \frac{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i Y_i}}{\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i}^2} = \frac{\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i Y_i}}{\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X_i}^2} = \frac{E(X_i Y_i)}{E\left(X_i{}^2\right)} = \frac{\frac{1}{N}\sum_{k=1}^{N} x_k y_k}{\frac{1}{N}\sum_{k=1}^{N} x_k{}^2} = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} x_k{}^2} = \alpha.$$

# What you need to remember

- Prediction for $y_k$ based on OLS regression in population is $\alpha x_k$, with $\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$.
- We would like to compute $\alpha$ but we cannot because we do not observe the $y_k$s of everybody in the population.
- => we randomly draw $n$ units with replacement from the population, and measure the dependent and the independent variable of those $n$ units.
- For every $i$ between 1 and $n$, $Y_i$ and $X_i$ = value of dependent and independent variables of the $i$th unit we randomly select.
- Given that $\alpha$ is value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$, we use $\hat{\alpha}$, the value of $a$ that minimizes $\sum_{i=1}^{n}(Y_i - aX_i)^2$ to estimate $\alpha$.
- We have $\hat{\alpha} = \frac{\sum_{i=1}^{n} \boldsymbol{X_i Y_i}}{\sum_{i=1}^{n} \boldsymbol{{X_i}^2}}$.
- Law of large numbers implies that $\lim_{n \to +\infty} \hat{\alpha} = \alpha$.
- When the sample we randomly draw gets large, $\hat{\alpha}$, the sample coefficient of the regression, gets close to $\alpha$, the population coefficient.
- Therefore, $\hat{\alpha}$ is a good proxy for $\alpha$ when the sample size is large enough.

# Roadmap

1. The OLS univariate linear regression function.
2. Estimating the OLS univariate linear regression function.
3. OLS univariate linear regression in practice.

# How Gmail uses univariate linear regression (1/2)

- Gmail would like to predict $y_k$, a variable equal to 1 if email $k$ is a spam and 0 otherwise.
- To do so they use a variable $x_k$, a variable equal to 1 if the word "free" appears in the email and 0 otherwise.
- $x_k$ is easy to measure (a computer can do it automatically, by doing a search of "free" in the email), but $y_k$ is hard to measure: only a human can know for sure whether an email is a spam or not. => they cannot observe $y_k$ for all emails reaching Gmail.
- To make good predictions, they would like to compute, $\alpha$, the value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$, and then use $\alpha x_k$ to predict $y_k$. $\alpha x_k$: best univariate linear prediction of $y_k$ given $x_k$.
- Issue: $\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$ => they cannot compute it unless they observe all the $y_k$s, but that would be very costly to do (a human has to read all the emails reaching Gmail accounts), plus once it's done we no longer need to predict the $y_k$s because we know them.

# How Gmail uses univariate linear regression (2/2)

- Instead Gmail can draw a random sample of, say, 5000 emails, ask humans to read them and determine whether they are spams or not.
- For every $i$ between 1 and $5000$, let $Y_i$ denote whether the $i$th randomly drawn email is a spam or not, and let $X_i$ denote whether the $i$th randomly drawn email has the word free in it.
- Then, people in Gmail can compute $\hat{\alpha} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}$.
- For all the emails they have not randomly drawn, and for which they do not observe $y_k$, they can use $\hat{\alpha} x_k$ as their prediction of whether the email is a spam or not.
- Because their random sample of emails is large, $\hat{\alpha}$ should be close to $\alpha$, and therefore $\hat{\alpha} x_k$ should be close to $\alpha x_k$, the best univariate linear prediction of $y_k$ given $x_k$.

# How banks use univariate linear regression (1/2)

- A bank would like to predict $y_k$, a variable equal to the amount that a person applying in April 2018 for a one-year loan will fail to reimburse in April 2019 when her loan expires.
- To do so they use a variable $x_k$, equal to the FICO score of that applicant.
- $x_k$ is easy to measure (the bank has access to the FICO score of all applicants), but $y_k$ is impossible to measure today: it's only in April 2019 that the bank will know the amount the applicant fails to reimburse.
- To make good predictions, they would like to compute $\alpha$, the value of $a$ that minimizes $\sum_{k=1}^{N}(y_k - ax_k)^2$, and then use $\alpha x_k$ to predict $y_k$. $\alpha x_k$: best univariate linear prediction of $y_k$ given $x_k$.
- Issue: $\alpha = \frac{\sum_{k=1}^{N} x_k y_k}{\sum_{k=1}^{N} {x_k}^2}$ => they cannot compute because they do not observe the $y_k$s.

# How banks uses univariate linear regression (2/2)

- Instead, the bank can use data on people who applied in April 2017 for a one-year loan. For those people, they know how much they failed to reimburse on their loan. Let's assume that the bank has 1000 applicants in April 2018, and 1000 applicants in April 2017.
- For every $i$ between 1 and $1000$, let $Y_i$ denote the amount that the $i$th April 2017 applicant failed to reimburse on her loan, and let $X_i$ denote the FICO score of that applicant.
- Then, people in the bank can compute $\hat{\alpha} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} {X_i}^2}$.
- For their April 2018 applicants, for which they do not observe $y_k$, they can use $\hat{\alpha} x_k$ as their prediction of the amount each applicant will fail to reimburse.
- Which condition should be satisfied to ensure $\hat{\alpha}$ is close to $\alpha$? Hint: look again at the Gmail example. There is one difference in the way we select the observations for which we measure $Y_i$ in the bank and in the Gmail example. Discuss this question with your neighbor for one minute.

# iClicker time

- Which condition should be satisfied to ensure $\hat{\alpha}$ is close to $\alpha$?

a) The April 2017 applicants should look very similar to the April 2018 applicants (e.g.: they should have similar FICO scores, etc.)

b) The number of April 2017 applicants should be close to the number of April 2018 applicants.

# April 2017 and 2018 applicants should look similar

- Previous section: $\hat{\alpha}$ converges towards $\alpha$ if sample of units for which we observe $Y_i$ randomly drawn from population.
- The bank cannot draw random sample of April 2018 applicants and observe today the amount this sample will fail to reimburse.
- Instead, can use the April 2017 applicants, for which they can both measure $Y_i$, amount that each applicant failed to reimburse, and $X_i$, FICO score.
- Then, can compute $\hat{\alpha}$, and for each April 2018 applicant they can use $\hat{\alpha}x_k$ as their prediction of $y_k$, the amount each applicant will fail to reimburse.
- If April 2017 applicants are "as good as" a random sample from combined population of April 2017 and April 2018 applicants, then all our theoretical results apply: $\hat{\alpha}$ should be close to $\alpha$, and our predictions should be good.
- To assume that April 2017 applicants are almost a random sample from the population of April 2017 and April 2018 applicants, April 2017 and April 2018 should look very similar. E.g.: should have similar FICO scores, demographics...
- => if two groups look similar, $\hat{\alpha}x_k$ should be good prediction of $y_k$ for 2018 applicants. Otherwise, we have to be careful.

# What you need to remember, and what's next

- In practice, there are many instances where we can measure the $y_k s$, the variable we do not observe for everyone, for a subsample of the population.
- We can use that subsample to compute $\hat{\alpha}$ , and then use $\hat{\alpha} x_k$ as our prediction of the $y_k s$ we do not observe.
- If that subsample is a random sample from the population (Gmail example), $\hat{\alpha} x_k$ should be close to $\alpha x_k$, best linear prediction for $y_k$.
- On the other hand, if that subsample is not a random sample from the population (bank example),$\hat{\alpha} x_k$ will be close to $\alpha x_k$ only if the subsample looks pretty similar to the entire population (almost a random sample).
- Even when we have a random sample, univariate linear regression might still not give great predictions.
- There are better prediction methods available. Next lectures: we see one of them.