Basics of regression analysis

Federico Tamagni

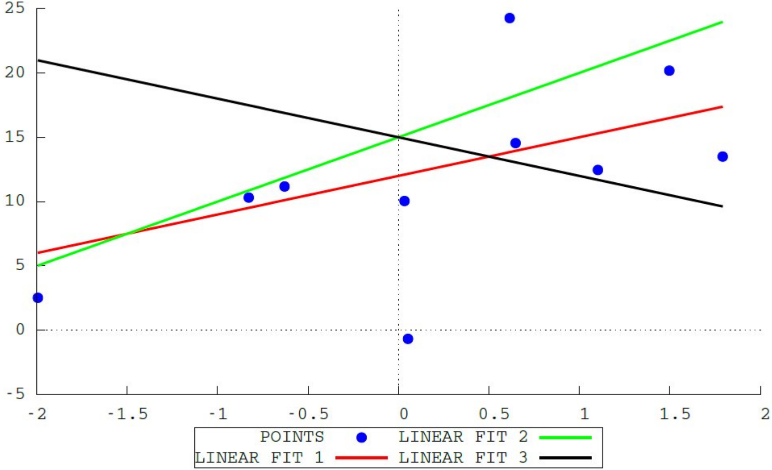
IE/LEM, Scuola Superiore Sant'Anna

Regression analysis

- Assume we have a collection of data on two economic quantities x and y for n individuals or unit of analysis, that is: $\{(x_i, y_i; i = 1, \dots, n\}$
- Suppose further that we would like to describe the relation between Y and X through the linear relation: $y = \beta_0 + \beta_1 x$

$\Longrightarrow$ How do we get the values of the parameters ?

Linear Fit Problem

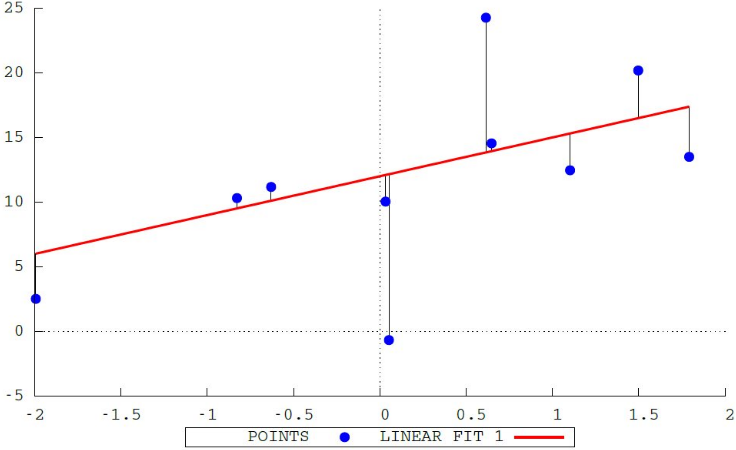


Linear Fit Problem

- Each line corresponds to a different slope and a different intercept
  - Linear Fit 1: $\hat{y} = 12 + 3 \times x$
  - Linear Fit 2: $\hat{y} = 15 + 5 \times x$
  - Linear Fit 1: $\hat{y} = 15 - 3 \times x$
- We need a way to assess which line is best description of the data
- A possible criterion to decide the best one is to start from the error we make by using each line instead of the data points.

$$u_i = \hat{y}_i - y_i \tag{1}$$

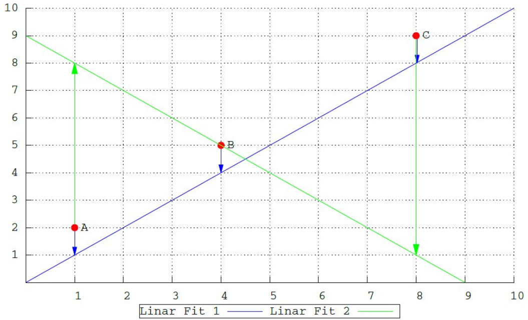
Linear Fit Problem



Black segments represent the error we make by substituting y_i with $\hat{y}_i$, that is $error_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i$. In this example is $y_i - 12 - 3x_i$.

Linear Fit Problem

- For each pair (x_i, y_i) you have an error u_i, so we can compute an *overall* error just by summing, that is $\sum_i u_i$
- Is it ok ? Consider an example with 2 possible linear fits for a set of 3 data points



Linear Fit Problem

- For each pair (x_i, y_i) you have an error u_i, so we can compute an *overall* error just by summing, that is $\sum_i u_i$
- Is it ok ? Consider an example with 2 possible linear fits for a set of 3 data points
  - ✓ Linear Fit 1: errors are (1,1,1), so Sum of errors=3
  - ✓ Linear Fit 2: errors are (-6,0,8), so Sum of errors=2

$\Longrightarrow$ Sum is smaller when we actually make larger errors for 2 out of 3 points !!!

Linear Fit Problem

- For each pair (x_i, y_i) you have an error u_i, so we can compute an *overall* error just by summing, that is $\sum_i u_i$
- Is it ok ? Consider an example with 2 possible linear fits for a set of 3 data points
  - ✓ Linear Fit 1: errors are (1,1,1), so Sum of errors=3
  - ✓ Linear Fit 2: errors are (-6,0,8), so Sum of errors=2

$\Longrightarrow$ Sum is smaller when we actually make larger errors for 2 out of 3 points !!!

Linear Fit Problem

- To balance-out the effect of cancellations between positive and negative errors, we define the Sum of Squared Residuals

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- We take this as the overall size of the mistake made when using each linear fit

$\Longrightarrow$ A natural criterion to estimate the parameter is to find values of β_0 and β_1 that minimize the SSR: this is called Ordinary Least Square (OLS) method

The OLS criterion for Linear Fitting

Formally the problem is

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{\text{MIN}} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The two necessary conditions to identify the solution read

$$\begin{cases} \frac{\partial \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0 \end{cases} \qquad \begin{cases} -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

with solutions

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{COV(x,y)}{VAR(x)} \end{cases} .$$

The OLS criterion for Linear Fitting

- Once we have the numbers $\hat{\beta}_0$ and $\hat{\beta}_1$ for a given data set, we write the OLS fitted line as a function of x:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The OLS fitted line allows us to predict y for any (sensible) value of x.
- The intercept, $\hat{\beta}_0$, is the predicted y when $x = 0$. (The prediction is usually meaningless if $x = 0$ is not possible.)
- The slope, $\hat{\beta}_1$, allows us to predict changes in y for any (reasonable) change in x:

$$\Delta\hat{y} = \hat{\beta}_1 \Delta x$$

- If $\Delta x = 1$, so that x increases by one unit, then $\Delta\hat{y} = \hat{\beta}_1$.

Algebraic properties of OLS

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{\text{MIN}} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Necessary conditions to identify the solution are

$$\begin{cases} \frac{\partial \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0 \end{cases} \qquad \begin{cases} -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ -2\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

$$\begin{cases} \sum_i \hat{u}_i = 0 \\ \sum_i \hat{u}_i x_i = 0 \end{cases}$$

AP1. OLS residuals always add up to 0

AP2. Sample covariance between residuals $\hat{u}_i$ and x_i is 0.

Why we like OLS ?

- Suppose to have observations on a population of individuals (even if most often we work with samples)
- We have data (Y_i, X_i), where this time X is a set of characteristics (not just one variable)
- We are interested in understanding to what extent knowledge of Xs helps to characterize Y, or, similarly, to explain or predict Y on the basis of the Xs. That is, we are interested in some function of Y conditional on the Xs
    - Y is the dependent variable
    - The Xs are called covariates (aka regressors or explanatory variables)

$\Longrightarrow$ We spend next slides to see in which sense OLS are "good" in achieving this aim

Why we like OLS ?

- Suppose to have observations on a population of individuals (even if most often we work with samples)
- We have data (Y_i, X_i), where this time X is a set of characteristics (not just one variable)
- We are interested in understanding to what extent knowledge of Xs helps to characterize Y, or, similarly, to explain or predict Y on the basis of the Xs. That is, we are interested in some function of Y conditional on the Xs
  - Y is the dependent variable
  - The Xs are called covariates (aka regressors or explanatory variables)

$\Longrightarrow$ We spend next slides to see in which sense OLS are "good" in achieving this aim

Why we like OLS ?

Theorem (CEF decomposition properties)

$$Y_i = E[y_i|X_i] + \epsilon_i \quad ,$$

where

- $E[\epsilon_i|X_i] = 0$;
- ϵ_i *is uncorrelated with any function of* X_i.

$\Longrightarrow$ Any variable y_i can be decomposed into two pieces:

- A piece which is orthogonal to any function of the Xs
- A piece that is explained by the Xs, captured by the CEF

- $E[Y_i \mid X_i] = \int dt\, t\, f_y(t \mid X_i = x)$, with f_y a conditional density, is called the Conditional Expectation Function (CEF)

Why we like OLS ?

Theorem (CEF decomposition properties)

$$Y_i = E[y_i|X_i] + \epsilon_i \quad ,$$

where

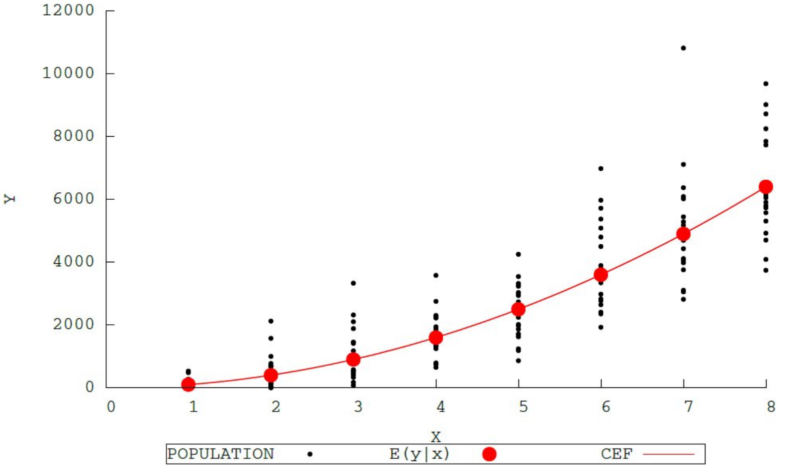
- $E[\epsilon_i|X_i] = 0$;
- ϵ_i *is uncorrelated with any function of* X_i.

$\Longrightarrow$ Any variable y_i can be decomposed into two pieces:

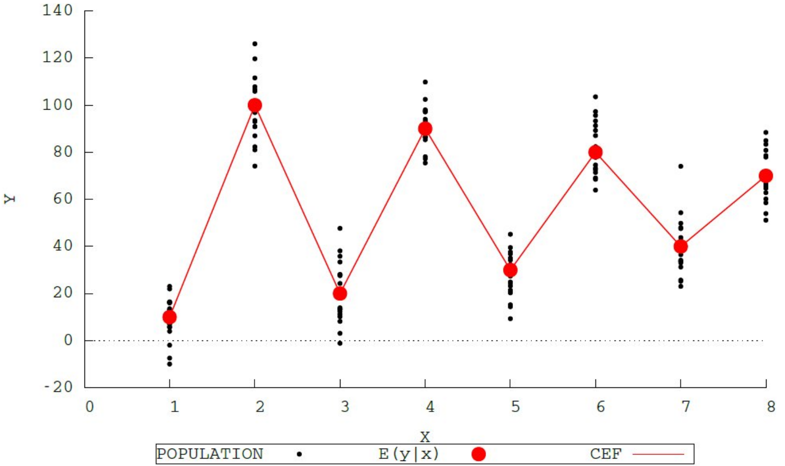
- A piece which is orthogonal to any function of the Xs
- A piece that is explained by the Xs, captured by the CEF

- $E\left[Y_i \mid X_i\right] = \int dt\; t\, f_y(t \mid X_i = x)$, with f_y a conditional density, is called the Conditional Expectation Function (CEF)

CEF example 1



CEF example 2



Why we like OLS ?

Theorem (CEF prediction property)

Let $m(X_i)$ be any function of X_i. Then

$$E[y_i|X_i] = \underset{m(X_i)}{arg\ min}\ E\left[(y_i - m(X_i))^2\right]$$

so the CEF is the MMSE predictor of y_i given X_i.

$\Longrightarrow$ The CEF is the best predictor of Y given the Xs, among *all possible* functions of the Xs (in MSE terms)

✓ The CEF represents a "precise" way to characterize the relationship between Y and the Xs, if we ask "how can we explain or predict Y, based on info about the Xs ?"

Why we like OLS ?

Theorem (CEF prediction property)

Let $m(X_i)$ be any function of X_i. Then

$$E[y_i|X_i] = \underset{m(X_i)}{arg\ min}\ \ E\left[(y_i - m(X_i))^2\right]$$

so the CEF is the MMSE predictor of y_i given X_i.

$\Longrightarrow$ The CEF is the best predictor of Y given the Xs, among *all possible* functions of the Xs (in MSE terms)

✓ The CEF represents a "precise" way to characterize the relationship between Y and the Xs, if we ask "how can we explain or predict Y, based on info about the Xs ?"

Why we like OLS ? The CEF-OLS link

Theorem (the regression-CEF theorem)

The function $X_i'\beta$ provides the MMSE linear approximation to $E[y_i|X_i]$, that is

$$\beta = \underset{b}{arg\ min}\ \ E\left\{(E[y_i|X_i] - X_i'b)^2\right\}\ . \tag{2}$$

Proof.

Write

$$\begin{aligned}(y_i - X_i'b)^2 &= (y_i - E[y_i|X_i])^2 + (E[y_i|X_i] - X_i'b)^2 \\ &\quad + 2(y_i - E[y_i|X_i])(E[y_i|X_i] - X_i'b)\ .\end{aligned}$$

The first term does not involve b and the last one has expectation zero by the CEF-decomposition property.

$\Longrightarrow$ Problem (2) has same solution as $\text{MIN}_b\ E\left[(Y_i - X_i'b)^2\right]$, which is exactly what we do in OLS !!!

✓ The linear regression function $X'b$ that we get from OLS is a good approximation (the best in MMSE terms) of the CEF

Why we like OLS ? The CEF-OLS link

Theorem (the regression-CEF theorem)

The function $X_i'\beta$ provides the MMSE linear approximation to $E[y_i|X_i]$, that is

$$\beta = \underset{b}{arg\ min} \quad E\left\{(E[y_i|X_i] - X_i'b)^2\right\} \quad . \tag{2}$$

Proof.

Write

$$(y_i - X_i'b)^2 = (y_i - E[y_i|X_i])^2 + (E[y_i|X_i] - X_i'b)^2 \\ + 2(y_i - E[y_i|X_i])(E[y_i|X_i] - X_i'b) \quad .$$

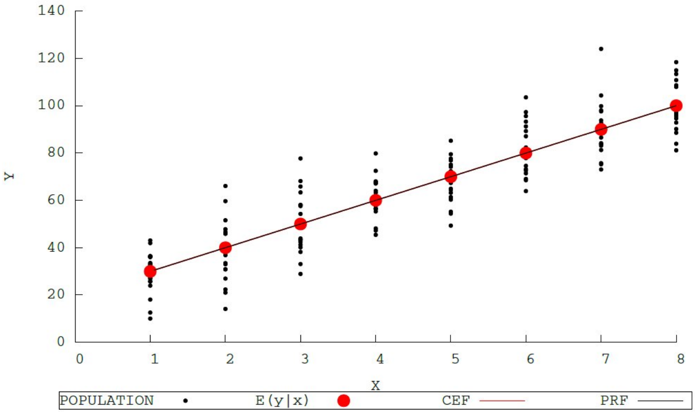
The first term does not involve b and the last one has expectation zero by the CEF-decomposition property.

$\Longrightarrow$ Problem (2) has same solution as $\text{MIN}_b\ E\left[(Y_i - X_i'b)^2\right]$, which is exactly what we do in OLS !!!

✓ The linear regression function $X'b$ that we get from OLS is a good approximation (the best in MMSE terms) of the CEF

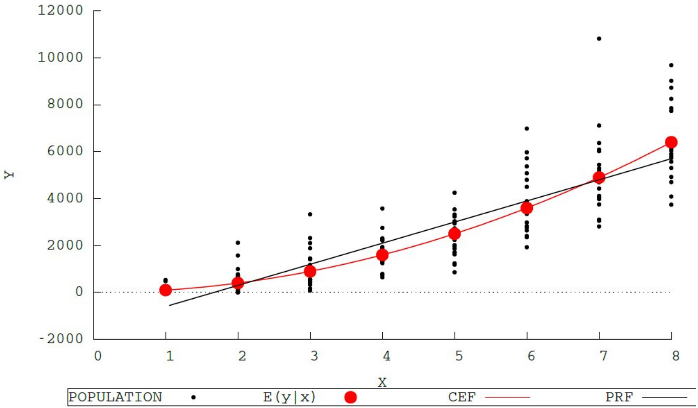
How good is the OLS-Population Regression Function ?

If the CEF is linear, then the (linear) Population Regression Function is exactly the CEF

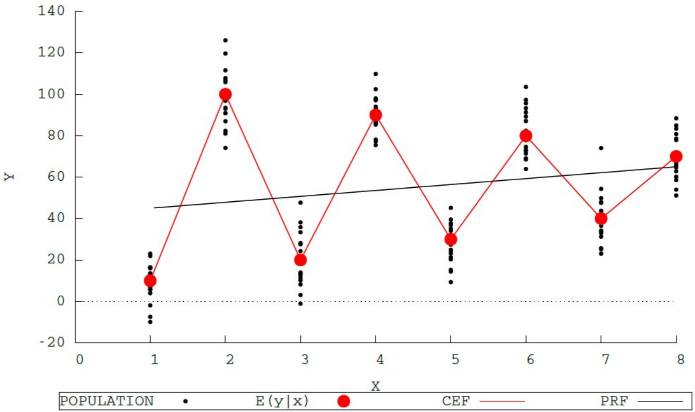


How good is the OLS-PRF ?

In other situations, it is the best we can do (in MMSE terms), but not always satisfactory



In other situations, it is the best we can do (in MMSE terms), but not always satisfactory



OLS Regression recap and properties

- We have learnt that each y_i can be expressed as

$$y_i = E(y|x_i) + u_i \quad ,$$

  where the error term u_i captures how much we are distant from the CEF

- What is in the error term ?
  - omitted factors, due to a wrong idea or theory about what we should consider as predictors (or determinants) of y
  - omitted factors due to lacking data on some Xs that we would like to include
  - wrong functional form

OLS Regression recap and properties

- We have learnt that each y_i can be expressed as

$$y_i = E(y|x_i) + u_i \quad ,$$

where the error term u_i captures how much we are distant from the CEF

- What is in the error term ?
  - omitted factors, due to a wrong idea or theory about what we should consider as predictors (or determinants) of y
  - omitted factors due to lacking data on some Xs that we would like to include
  - wrong functional form

- Crucial property is the Zero Conditional Mean (ZCM) property:

> ZCM. The average, or expected, value of u, conditional on x, is zero. Formally, $E(u|x) = 0$,

- ZCM is crucial (with some other assumptions) to show that the OLS estimates are unbiased
- ZCM is crucial to show consistency of OLS estimates: $\hat{\beta}$ converges in prob to β

OLS Regression recap and properties

- Crucial property is the Zero Conditional Mean (ZCM) property:

> ZCM. The average, or expected, value of u, conditional on x, is zero. Formally, $E(u|x) = 0$,

- ZCM is crucial (with some other assumptions) to show that the OLS estimates are unbiased
- ZCM is crucial to show consistency of OLS estimates: $\hat{\beta}$ converges in prob to β

OLS Regression recap and properties

- Do not forget that OLS confine the attention to the expected value of the conditional distribution of y give X
  - In general, one may be interested into other features of the same distribution, so we would need different techniques in that case
  - The conditional expected value might be particularly meaningless if y has a very skewed distribution, since in that case the mean of y says little
- Do not forget that the OLS weights a lot large errors (taking squares of the distance from linear fit line)
  - This means that few outliers can dramatically influence the estimates of the parameters
  - Often you either drop the outliers (if they are just few data-points) or look for different techniques that are less influenced by outliers

OLS Regression recap and properties

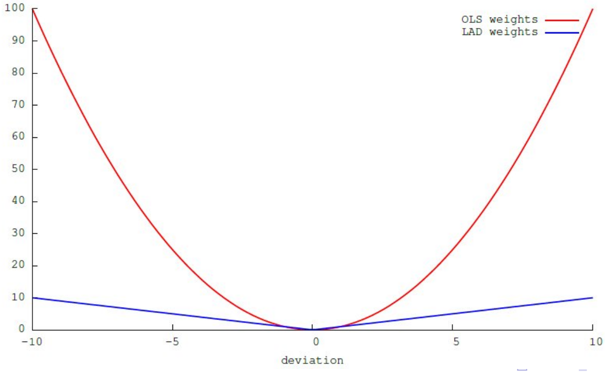
- Do not forget that OLS confine the attention to the expected value of the conditional distribution of y give X
  - In general, one may be interested into other features of the same distribution, so we would need different techniques in that case
  - The conditional expected value might be particularly meaningless if y has a very skewed distribution, since in that case the mean of y says little
- Do not forget that the OLS weights a lot large errors (taking squares of the distance from linear fit line)
  - This means that few outliers can dramatically influence the estimates of the parameters
  - Often you either drop the outliers (if they are just few data-points) or look for different techniques that are less influenced by outliers

OLS Regression recap and properties

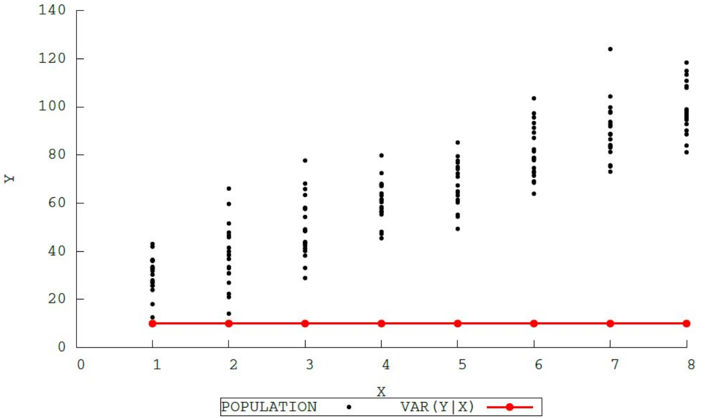
- One popular and easy correction for outliers is Least Absolute Deviation (LAD)

$$\underset{\hat\beta_0,\hat\beta_1}{\text{MIN}} \sum_i (y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 \qquad \underset{\hat\beta_0,\hat\beta_1}{\text{MIN}} \sum_i \left| y_i - \hat\beta_0 - \hat\beta_1 x_i \right|$$

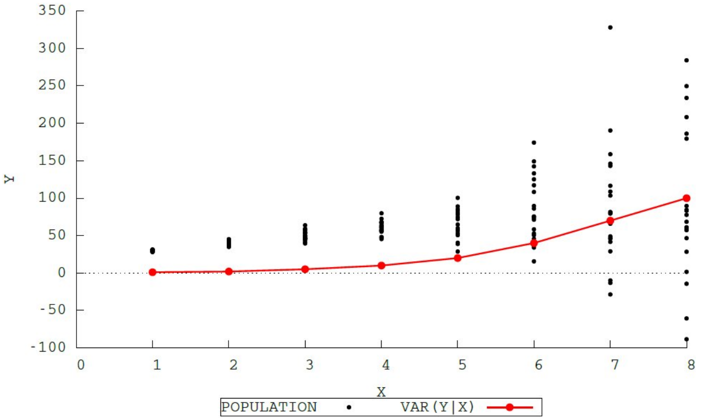
OLS estimation LAD estimation



- Another property worth noticing is HOMOSKEDASTICITY, meaning that the variance of the error (and thus of the part of Y not captured by the regression) is the same for every value of x: $\text{VAR}(u \mid x) = \sigma^2 > 0 \;, \forall x$



- HETEROSKEDASTICITY, instead, means that the variance of u (or of the part of Y not captured by the regression line) varies with the values of x



Regression analysis: final remarks

- This was a highly simplified presentation: many other problems remain in practical work
- The practice of econometrics is mostly to deal with real-world situation where OLS assumptions are difficult to maintain (e.g., recall sample-selection or endogeneity discussed in Gibrat's regression)
- Maximum-likelihood is a general alternative, flexible and able to also account for non-linearities: the idea is to assume a distribution for the errors, and then write the joint probability density and maximize it (e.g., recall above discussion about parametric density estimation, or the non-linear estimation of the scaling relationship between variance of growth and size)
- Nevertheless, it is useful as a benchmark: economists like to frame their research questions as "Is there an impact of a certain variable X on the value of an outcome value Y, and if so how strong is it ?"