Linear models

Gabriela Ciolek

Gabriela.ciolek@telecom-paristech.fr

Two variables

- y = explained variable, dependent
- x = explanatory variable, independent
- u = non-observed variable, error, disturbance

$$y = \beta_0 + \beta_1 x + u.$$

- β_0 = constant, intercept
- β_1 = partial effect = $\partial y / \partial x$ (slope parameter in the relationship between y and x holding the other factors in u fixed)

Example 1

- y = salary
- x = education
- u = labor force experience, productivity, tenure with current employer

$$wage = \beta_0 + \beta_1 educ + u.$$

- Ceteris paribus ? Explain link x-u

Example 2

- y = yield
- x = fertilizer
- u = quality of land, climat etc.

$$yield = \beta_0 + \beta_1 fertilizer + u,$$

- Partial effect, ceteris paribus. $\Delta yield = \beta_1 \Delta fertilizer.$
- Ceteris paribus ? Explain link x-u

Fundamental hypothesis

- Error variables have zero mean
- Conditional expectation = unconditional expectation
- X and u are uncorrelated

$$E(u|x) = E(u) = 0,$$

- Education (x) and tenure with current employer (u)

Population regression function

$$E(y|x) = \beta_0 + \beta_1 x$$

- E(y|x) is a linear function of x (the linearity means that a one-unit increase in x changes the expected value of y by ß1

Predictions

- Fitted value

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

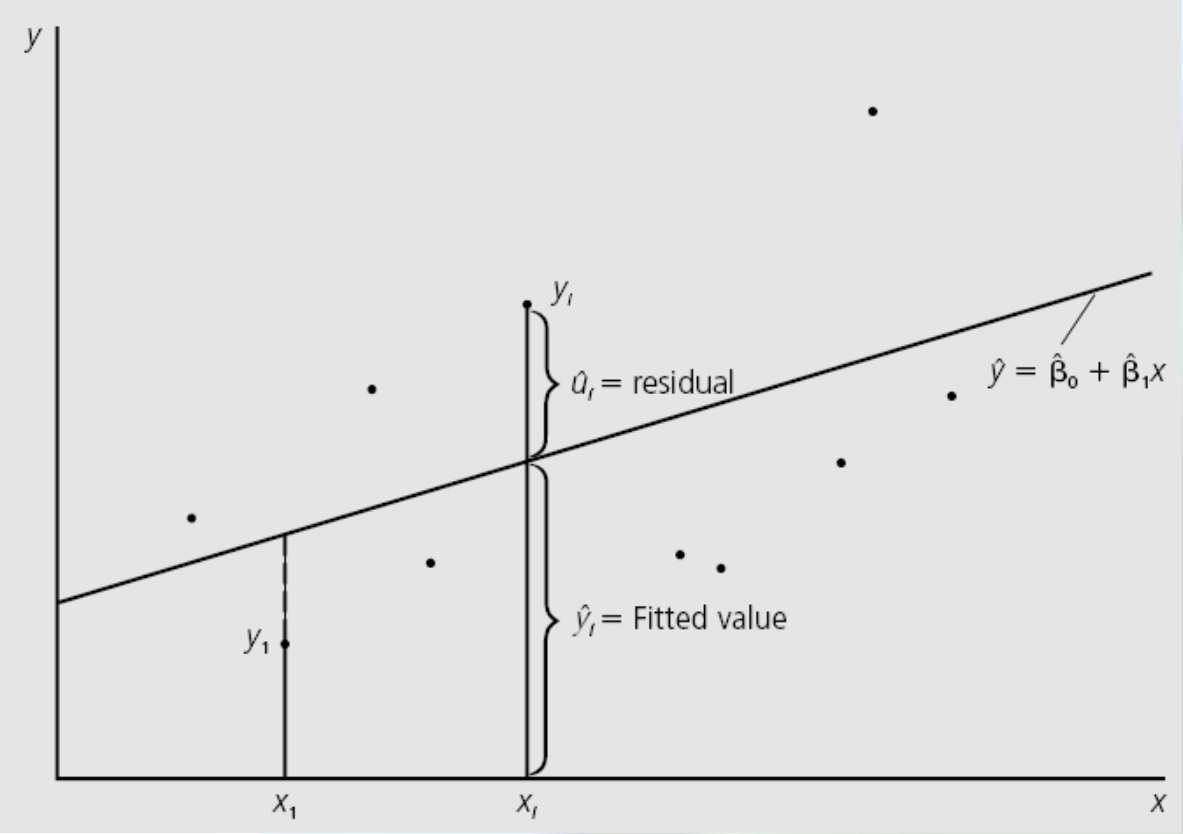
- Residual

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i.$$

- Sum of squared residuals

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2,$$

Graphical illustration



K explanatory variables

- Multiple linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + u,$$

- Fundamnetal hypothesis

$$\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0.$$

(all factors in the unobserved error term be uncorrelated with the explanatory variables)

Obtaining the OLS Estimates

- We minimize the sum of squared estimators

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2.$$

- Matrix notation:
- Min $(y - x\beta)'(y - x\beta)$

- First order conditions (concavity criteria)
- $(y - xb)' x = 0$
- Remark: we assume $x_0 = 1$

Matrix notation

- $X'(y - Xb) = 0$
- $X'y - X'Xb = 0$
- $b = (X'X)^{-1}X'y$
- Problem if $X'X \approx 0$: multicollinearity

(we call this problem collinearity: it looks like we have p different predictor variables, but really some of them are linear combinations of the others, so they don't add any information)

- $X'X$ = "variance"
- $X'y$ = "covariance"

Unbiased estimators

- The OLS estimators are unbiased estimators of the population parameters
- $E(b) = E(X'X)^{-1}X'y = E(X'X)^{-1}X'(X\beta+u) = \beta$

- Fundamental hypothesis
- $E(u \mid X) = 0$

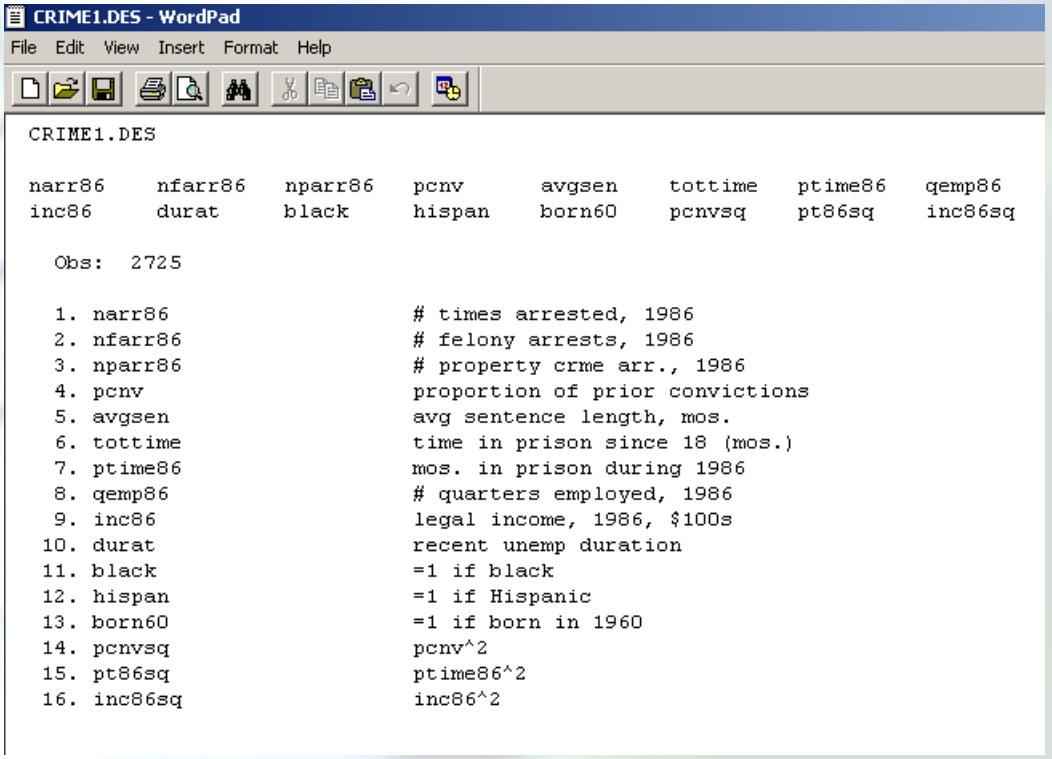
Data analysis with Matlab

Loading the data

```
load CRIME1.raw
```

Description of the data

- CRIME1.des

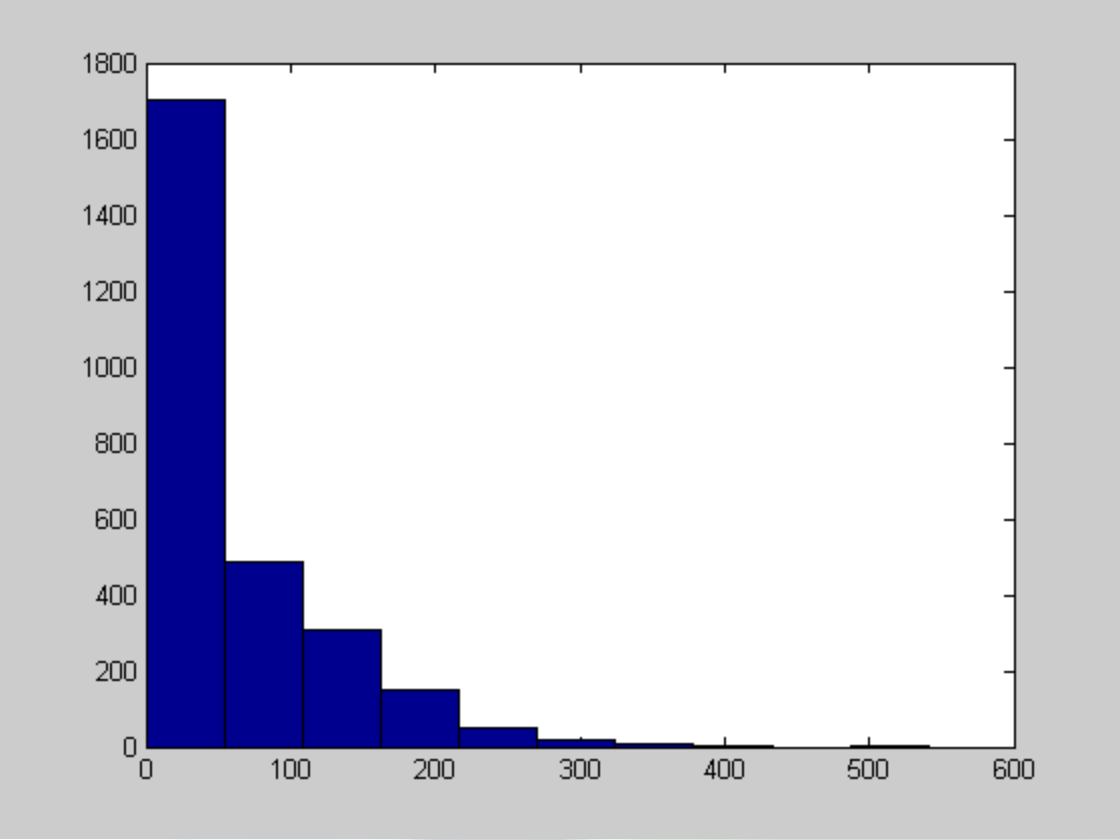


Exercise

- Draw a histogram of incomes in dollars in 1986.

Histogram

hist(crime1(:,9))

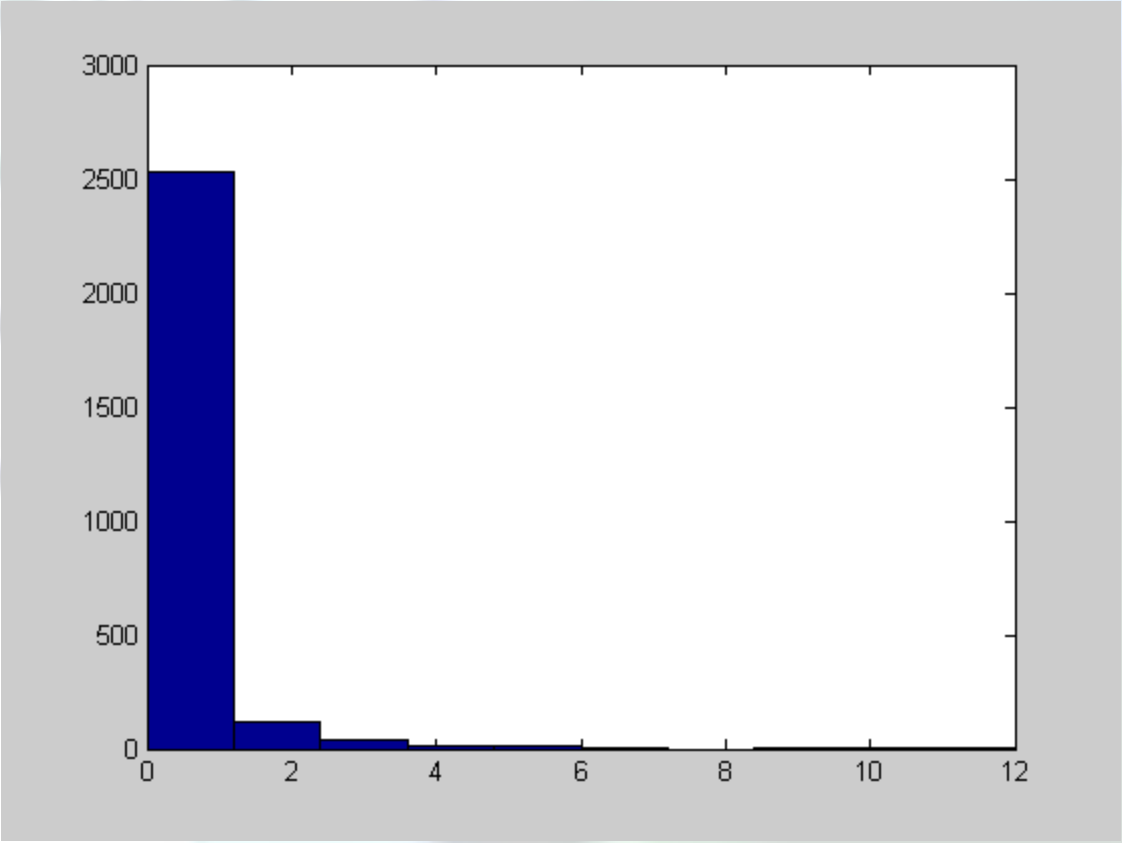


Exercise

- Draw a histogram of number of times when an individual was arrested in 1986.

Histograme

hist(crime1(:,1))



Exercise

Calculate:

- expectation
- variance
- correlation

between the number of times when the individual was arrested and other variables.

Descriptive statistics

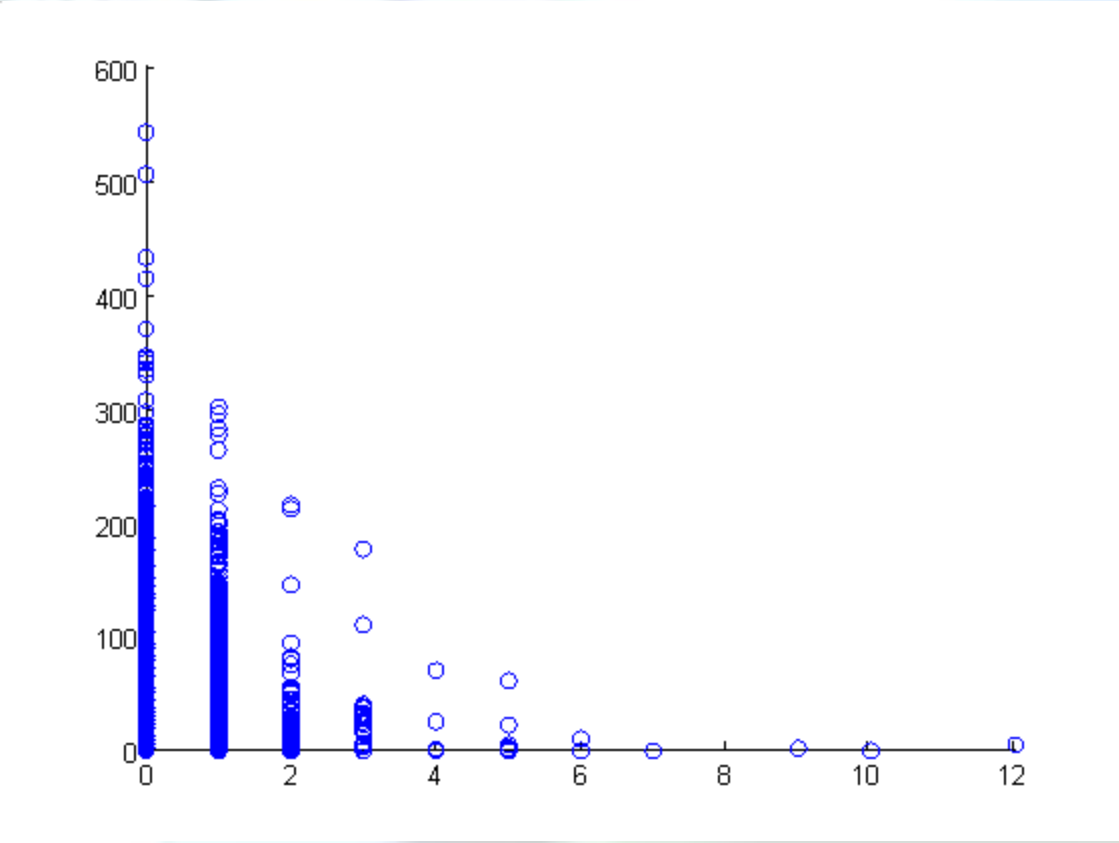
```
mean(crime1)’
mean(crime1(:,1:13))’
std(crime1(:,1:13))’
min(crime1)’
max(crime1)’
cov(crime1)’
corrcoef(crime1(:,1:13))'
```

Exercise

Draw the cloud of points (X,Y) with X the income and Y the numer of arrests

Scatterplots

```
scatter(crime1(:,1),crime1(:,9))
```



Analysis of the data

Import of the data

load WAGE1.raw

Description of the data

- WAGE1.des

WAGE1.DES - WordPad

File Edit View Insert Format Help

```
WAGE1.DES

wage       educ       exper      tenure     nonwhite   female     married    numdep
smsa       northcen   south      west       construc   ndurman    trcommpu   trade
services   profserv   profocc    clerocc    servocc    lwage      expersq    tenursq

  Obs:   526

  1. wage                   average hourly earnings
  2. educ                   years of education
  3. exper                  years potential experience
  4. tenure                 years with current employer
  5. nonwhite               =1 if nonwhite
  6. female                 =1 if female
  7. married                =1 if married
  8. numdep                 number of dependents
  9. smsa                   =1 if live in SMSA
 10. northcen               =1 if live in north central U.S
 11. south                  =1 if live in southern region
 12. west                   =1 if live in western region
 13. construc               =1 if work in construc. indus.
 14. ndurman                =1 if in nondur. manuf. indus.
 15. trcommpu               =1 if in trans, commun, pub ut
 16. trade                  =1 if in wholesale or retail
 17. services               =1 if in services indus.
 18. profserv               =1 if in prof. serv. indus.
 19. profocc                =1 if in profess. occupation
 20. clerocc                =1 if in clerical occupation
 21. servocc                =1 if in service occupation
 22. lwage                  log(wage)
 23. expersq                exper^2
 24. tenursq                tenure^2
```

Comparison of salaries between men and women

Calculate:

- The average hourly wage for two sexes
- Average hourly wage for each sex separately
- Histograms of the salaries for each sex
- Descriptive statistics for other variables for each gender (expected value, variance, etc.)

Data selection

- Women
- Number

```
s=(wage1(:,6)==1);
sum(s)
wage1f=wage1(s,:)
mean(wage1(s,1:21))'
```

Question

Is there a wage discrimination against women?

Response

We can not say based on the histograms of salaries of women and men because the other variables are not fixed.

It is possible that women can be les skilled so can be less paid, etc.

Explain wage | educ, exper, tenure

```
load WAGE1.raw
y=wage1(:,1);
[n,k]=size(wage1)
X=[ones(n,1),wage1(:,[2,3,4])];
[n,k]=size(X)
```

Estimation of parameters

Formula

$$beta=inv(X'*X)*X'*y$$

Calculate the residuals, variance of residuals and standard deviation of the estimator

u=y-X*beta
sig2=u'*u/(n-4) because 3 variables+ intercept
std=sqrt(diag(sig2*inv(X'*X)))

Exercise

What will happen if:

- One will multiply the salairy by 1000?
- One will multiply one explanatory variable by 1000?
- The coefficients will change? Are they robust to change unit?

Exercise

- Draw a histogram of residuals u.
- What are the proprieties of the distribution?

Exercise

- Do a regression when taking the log of salary, calculate the paremeter.
- Draw a new histogram for residuals.
- Proprieties?
- Detect the outliers, remove them and recalculate beta.

Exercise

Find observations such as u > 2.5

Remove those observations

```
I = find(u>=2.5)
s=(u<=2.5)
sum(s)
X=X(s,:);
y=y(s,:);
```