Multiple Linear Regression

Introduction to Statistical Learning
Bachelor in *Global Governance*
University of Rome - Tor Vergata

Marco Stefanucci
Department of Economics and Finance
University of Rome - Tor Vergata
marco.stefanucci@uniroma2.it

Introduction

- Regression models are used to describe how one or perhaps a few response variables depend on other explanatory variables.
- The idea of regression is at the core of much statistical modelling, because the question *what happens to y when x varies?* is central to many investigations.
- It is often required to predict or control future responses by changing the other variables, or to gain an understanding of the relation between them.
- There is usually a single response, treated as random. Often there are many explanatory variables, which are treated as non-stochastic.

If we denote the response by y and the explanatory variables by x, our concern is how changes in x affect y. Given (y_j, x_j) for $j = 1, \ldots, n$, in the previous lecture we fitted the straight-line regression model $y_j = \beta_1 + \beta_2 x_j + \epsilon_j$ for $j = 1, \ldots, n$

An immediate generalization is to increase the covariates,

$$y_j = \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots, \beta_p x_{jp} + \epsilon_j = x_j^t \beta + \epsilon_j \quad j = 1, \ldots, n$$

where $x_j^t = (x_{j1}, \ldots, x_{jp})$ is a $1 \times p$ vector of covariates associated with the jth response, β is a $p \times 1$ vector of unknown parameters and ϵ_j is an unobserved error accounting for the discrepancy between the observed response y_j and $x_j^t \beta$.

Warning: In these slides we simplify notation by using y to represent both the response variable and the value it takes

Matrix notation

Set

$$x_j = \begin{bmatrix} x_{j1} \\ \vdots \\ x_{jr} \\ \vdots \\ x_{jp} \end{bmatrix} \quad j = 1, \ldots, n, \quad \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_r \\ \vdots \\ \beta_p \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1^t \\ \vdots \\ x_j^t \\ \vdots \\ x_n^t \end{bmatrix} = \begin{bmatrix} x_{11} & \ldots & x_{1r} & \ldots & x_{1p} \\ \vdots & \vdots & \vdots & \vdots & \\ x_{j1} & \ldots & x_{jr} & \ldots & x_{jp} \\ \vdots & \vdots & \vdots & \vdots & \\ x_{n1} & \ldots & x_{nr} & \ldots & x_{np} \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_j \\ \vdots \\ \epsilon_n \end{bmatrix},$$

the linear regression model with *design matrix* X can be written as

$$y = X\beta + \epsilon$$

In extended form:

$$\begin{array}{ccccc}
y_1 = \beta_1 \cdot x_{11} + & \dots & \beta_r \cdot x_{1r} + & \dots & +\beta_p \cdot x_{1p} + \epsilon_1 \\
\vdots & \vdots & \vdots & \vdots & \\
y_j = \beta_1 \cdot x_{j1} + & \dots & \beta_r \cdot x_{jr} + & \dots & +\beta_p \cdot x_{jp} + \epsilon_j \\
\vdots & \vdots & \vdots & \vdots & \\
y_n = \beta_1 \cdot x_{n1} + & \dots & \beta_r \cdot x_{nr} + & \dots & +\beta_p \cdot x_{np} + \epsilon_n
\end{array}$$

With $x_1^t = [1, \dots, 1]$, so that:

$$\begin{array}{ccccc}
y_1 = \beta_1 + & \dots & \beta_r \cdot x_{1r} + & \dots & +\beta_p \cdot x_{1p} + \epsilon_1 \\
\vdots & \vdots & \vdots & \vdots & \\
y_j = \beta_1 + & \dots & \beta_r \cdot x_{jr} + & \dots & +\beta_p \cdot x_{jp} + \epsilon_j \\
\vdots & \vdots & \vdots & \vdots & \\
y_n = \beta_1 + & \dots & \beta_r \cdot x_{nr} + & \dots & +\beta_p \cdot x_{np} + \epsilon_n
\end{array}$$

Straight-line regression in matrix notation

For the straight line regression model $y_j = \beta_1 + \beta_2 x_j + \epsilon_j$ for $j = 1, \ldots, n$, the matrix form of the model is

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

so X is an $n \times 2$ matrix and β a 2×1 vector of parameters.

Least square estimates

The least square estimate of β is obtained by the value that minimizes the *sum of squares*

$$SS(\beta) = \sum_{j=1}^{n} (y_j - x_j^t \beta)^2 = (y - X\beta)^t (y - X\beta)$$

We obtain the least square estimate of β by solving the equations

$$\begin{aligned}
\frac{\partial SS(\beta)}{\partial \beta_1} &= -2\sum_{j=1}^{n} (y_j - x_j^t \beta) x_{j1} = 0 \\
&\vdots \\
\frac{\partial SS(\beta)}{\partial \beta_r} &= -2\sum_{j=1}^{n} (y_j - x_j^t \beta) x_{jr} = 0 \\
&\vdots \\
\frac{\partial SS(\beta)}{\partial \beta_p} &= -2\sum_{j=1}^{n} (y_j - x_j^t \beta) x_{jp} = 0
\end{aligned}$$

In matrix form these amount to the equations

$$(y - X\beta)^t X = (0, \dots 0)$$

that is

$$X^t(y - X\beta) = 0$$

which imply that the estimate satisfies

$$X^t y = X^t X\beta$$

Provided the $p \times p$ $X^t X$ is of full rank

$$\hat{\beta} = (X^t X)^{-1} X^t y$$

is the system solution.

In simple cases it is possible to have analytical expressions for the least square estimates. For example in the straight-line regression model the X matrix of the representation $y = X\beta + \epsilon$ is

$$X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

Then we have that

$$X^t X = \begin{pmatrix} n & \sum_{j=1}^n x_j \\ \sum_{j=1}^n x_j & \sum_{j=1}^n x_j^2 \end{pmatrix} \quad X^t y = \begin{pmatrix} \sum_{j=1}^n y_j \\ \sum_{j=1}^n x_j y_j \end{pmatrix}$$

After some algebra we obtain

$$\hat{\beta} = \begin{pmatrix} \bar{y} - \bar{x}\, s_{xy}/s_x \\ s_{xy}/s_x \end{pmatrix}$$

FITTED VALUES AND RESIDUALS

The sum of squares $SS(\beta)$ plays a central role. Its minimum value

$$SS(\hat{\beta}) = \sum_{j=1}^{n} (y_j - x_j^t \hat{\beta})^2 = (y - X\hat{\beta})^2 (y - \hat{\beta})$$

is called **residual sum of squares**. It is the squared discrepancy between the observations y and the **fitted values** $\hat{y} = X\hat{\beta}$.

The vector $\hat{y} = X\hat{\beta}$ is the linear combination of the columns of X that minimizes the squared distance with the data y.

Note that

$$\hat{y} = X\hat{\beta} = X(X^tX)^{-1}X^ty = Hy$$

The matrix $H = X(X^tX)^{-1}X^t$ is called *hat* matrix or *projection matrix*

The unobservable error $\epsilon_j = y_j - x_j^t \beta$ is estimated by the jth **residual**

$$e_j = y_j - x_j^t \hat{\beta}$$

In vector terms,

$$e = y - x\hat{\beta} = y - Hy = (I - H)y$$

where I is the $n \times n$ identity matrix.

Assuming that $E(\epsilon_j) = 0$ and $Var(\epsilon_j) = \sigma^2 = E(\epsilon_j^2)$ for $j = 1, \ldots, n$ we can estimate σ^2 with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} e_j^2 = \frac{e^t e}{n} = \frac{(y - \hat{y})^t (y - \hat{y})}{n}$$

Two groups comparison

Suppose that the response variable y has been observed on two groups of observations of size n_1 and n_2. Let y_{1j} for $j = 1, \dots n_1$ be the observations of the first group and let y_{2j} for $j = 1 \dots n_2$ be the observation of the second group.

Let β_1 and $\beta_1 + \beta_2$ be the means of the variable y in the two groups. Hence

$$\begin{aligned} y_{1j} &= \beta_1 + \epsilon_{1j} \quad j = 1, \dots, n_1 \\ y_{2j} &= \beta_1 + \beta_2 + \epsilon_{2j} \quad j = 1 \dots n_2 \end{aligned}$$

We can write the model for the two groups comparison in matrix notation $y = X\beta + \epsilon$ where

$$y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \end{pmatrix} \quad X = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \epsilon = \begin{pmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{2n_1} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}$$

For this model we have

$$\hat{\beta} = (X^t X)^{-1} X^t y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \end{pmatrix}$$

The two groups comparison can be extended to more than two groups

$$
\begin{array}{rcl}
y_{1j} & = & \beta_1 + \epsilon_{1j} \quad j = 1, \ldots, n_1 \\
y_{2j} & = & \beta_1 + \beta_2 + \epsilon_{2j} \quad j = 1 \ldots n_2 \\
\vdots & \vdots & \vdots \\
y_{kj} & = & \beta_1 + \beta_k + \epsilon_{kj} \quad j = 1 \ldots n_k
\end{array}
$$

Let $y_j = (y_{j1}, \ldots, y_{jn_j})^t$ for $j = 1, \ldots, k$.

$$
y = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} \quad X = \begin{pmatrix} 1_{n_1} & 0_{n_1} & \cdots & 0_{n_1} \\ 1_{n_2} & 1_{n_2} & \cdots & 0_{n_1} \\ \vdots & & \ddots & \\ 1_{n_k} & 0_{n_k} & \cdots & 1_{n_k} \end{pmatrix}
$$

$$
\hat{\beta} = (X^t X)^{-1} X^t y = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 - \bar{y}_1 \\ \vdots \\ \bar{y}_k - \bar{y}_1 \end{pmatrix}
$$

Some Important Questions

When we perform multiple linear regression, we usually are interested in answering a few important questions.

1. Is at least one of the predictors $x_1, x_2, \ldots, x_p$ useful in predicting the response?
2. Do all the predictors help to explain y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Is at least one of the predictors useful?

In the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. We test the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_1 : \text{at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the F-statistic:

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)}$$

Potential Problems

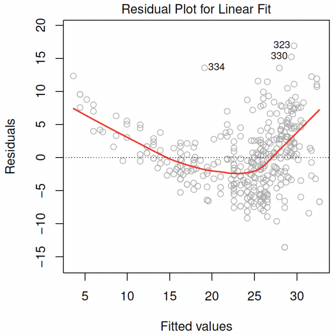
When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

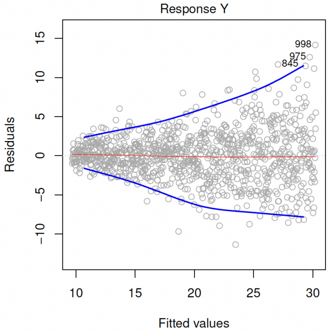
1. Non-linearity of the response-predictor relationships
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Residual Plots

Residual plots are a useful graphical tool for identifying non-linearity. Given a simple linear regression model, we can plot the residuals, $e_i = y_i - \hat{y}_i$, versus the predictor x_i. In the case of a multiple regression model, since there are multiple predictors, we instead plot the residuals versus the predicted (or fitted) values $\hat{y}_i$.

Ideally, the residual plot will show no discernible pattern. The presence of a pattern may indicate a problem with some aspect of the linear model.


Residual Plot for Linear Fit
Residuals
Fitted values
−15
−10
−5
0
5
10
15
20
5
10
15
20
25
30
323
330
334


Response Y
Residuals
Fitted values
−10
−5
0
5
10
15
10
15
20
25
30
998
975
845

Multicollinearity problem

- Multicollinearity (collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data.

- Let consider a $(n \times p)$ design matrix X; if $|cor(x_i, x_j)| = 1$ for $i \neq j$ and $i, j \in \{1, p\}$, then there is perfect collinearity and the product matrix $X^t X$ is not invertible.

Life cycle savings data

LifeCycleSavings is data set 5 with variables observed on 50 different countries. The variables are:

- **sr** aggregate personal savings,
- **pop15** % of population under 15,
- **pop75** % of population over 75,
- **dpi** real per-capita disposable income,
- **ddpi** % growth rate of dpi

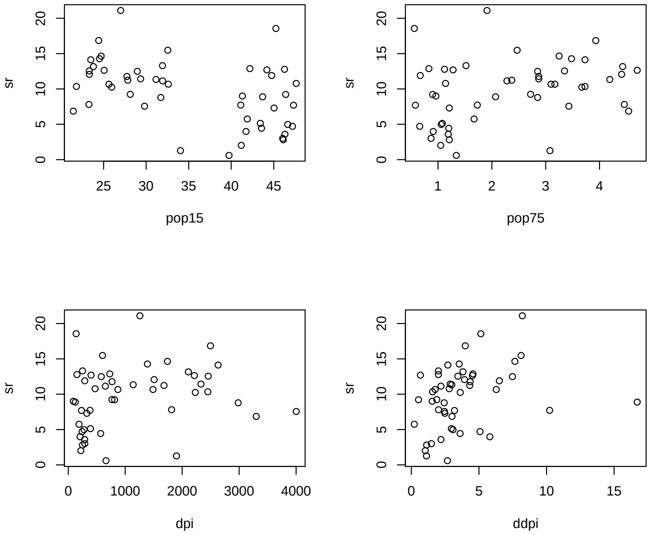
	sr	pop15	pop75	dpi	ddpi
Australia	11.43	29.35	2.87	2329.68	2.87
Austria	12.07	23.32	4.41	1507.99	3.93
Belgium	13.17	23.80	4.43	2108.47	3.82
Bolivia	5.75	41.89	1.67	189.13	0.22
Brazil	12.88	42.19	0.83	728.47	4.56
Canada	8.79	31.72	2.85	2982.88	2.43
Chile	0.60	39.74	1.34	662.86	2.67
China	11.90	44.75	0.67	289.52	6.51
Colombia	4.98	46.64	1.06	276.65	3.08
⋮	⋮	⋮	⋮	⋮	⋮
Turkey	5.13	43.42	1.08	389.66	2.96
Tunisia	2.81	46.12	1.21	249.87	1.13
United Kingdom	7.81	23.27	4.46	1813.93	2.01
United States	7.56	29.81	3.43	4001.89	2.45
Venezuela	9.22	46.40	0.90	813.39	0.53
Zambia	18.56	45.25	0.56	138.33	5.14
Jamaica	7.72	41.12	1.73	380.47	10.23
Uruguay	9.24	28.13	2.72	766.54	1.88
Libya	8.89	43.69	2.07	123.58	16.71
Malaysia	4.71	47.20	0.66	242.69	5.08

The data set is avaialble in R

```
> summary(LifeCycleSavings)
       sr             pop15           pop75            dpi
 Min.   : 0.600   Min.   :21.44   Min.   :0.560   Min.   :  88.94
 1st Qu.: 6.970   1st Qu.:26.21   1st Qu.:1.125   1st Qu.: 288.21
 Median :10.510   Median :32.58   Median :2.175   Median : 695.66
 Mean   : 9.671   Mean   :35.09   Mean   :2.293   Mean   :1106.76
 3rd Qu.:12.617   3rd Qu.:44.06   3rd Qu.:3.325   3rd Qu.:1795.62
 Max.   :21.100   Max.   :47.64   Max.   :4.700   Max.   :4001.89
      ddpi
 Min.   : 0.220
 1st Qu.: 2.002
 Median : 3.000
 Mean   : 3.758
 3rd Qu.: 4.478
 Max.   :16.710
```

Under the life-cycle savings hypothesis as developed by Franco Modigliani, the savings ratio (aggregate personal saving divided by disposable income) is explained by per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the percentage of the population over 75 years old.

The data are averaged over the decade 1960-1970 to remove the business cycle or other short-term fluctuations.


sr
0
5
10
15
20
25
30
35
40
45
pop15
sr
1
2
3
4
pop75
sr
0
1000
2000
3000
4000
dpi
sr
0
5
10
15
ddpi

In this case we might fit the model

$$y_j = \beta_1 + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \beta_5 x_{5j} + \epsilon_j$$

where y is the saving ratio and x_2, x_3, x_4 and x_5 are the variables pop15, pop75,dpi and ddpi. Looking the data we may expect a negative value for β_2 and a positive value for β_5 while the relationship between the saving ratio and the variables pop75 and dpi is not clear. The X matrix has dimension 50×5 and is

$$\begin{pmatrix} 1 & 29.35 & 2.87 & 2329.68 & 2.87 \\ 1 & 23.32 & 4.41 & 1507.99 & 3.93 \\ 1 & 23.80 & 4.43 & 2108.47 & 3.82 \\ 1 & 41.89 & 1.67 & 189.13 & 0.22 \\ 1 & 42.19 & 0.83 & 728.47 & 4.56 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 28.13 & 2.72 & 766.54 & 1.88 \\ 1 & 43.69 & 2.07 & 123.58 & 16.71 \\ 1 & 47.20 & 0.66 & 242.69 & 5.08 \end{pmatrix}$$

In R we can find the fitted values and the residual in the following way.

```
> m=lm(sr~pop15+pop75+dpi+ddpi,data=LifeCycleSavings)
> fitted(m)
     Australia        Austria        Belgium        Bolivia         Brazil
     10.566420      11.453614      10.951042       6.448319       9.327191
        Canada          Chile          China       Colombia     Costa Rica
      9.106892       8.842231       9.363964       6.431707       5.654922
       Denmark        Ecuador        Finland         France        Germany
     11.449761       5.995631      12.921086      10.164528      12.730699
        Greece      Guatamala       Honduras        Iceland          India
     13.786168       6.365284       6.989976       7.480582       8.491326
       Ireland          Italy          Japan          Korea     Luxembourg
      7.948869      12.353245      15.818514      10.086981      12.020807
         Malta         Norway    Netherlands    New Zealand      Nicaragua
     12.505090      11.121785      14.224454       8.384445       6.653603
        Panama       Paraguay           Peru    Philippines       Portugal
      7.734166       8.145759       6.160559       6.104992      13.258445
  South Africa South Rhodesia          Spain         Sweden    Switzerland
     10.656834      12.008566      12.441156      11.120283      11.643174
        Turkey        Tunisia United Kingdom  United States      Venezuela
      7.795682       5.627920      10.502413       8.671590       5.587482
        Zambia        Jamaica        Uruguay          Libya       Malaysia
      8.809086      10.738531      11.503827      11.719526       7.680869
```

```
> residuals(m)
     Australia        Austria        Belgium        Bolivia         Brazil
     0.8635798      0.6163860      2.2189579     -0.6983191      3.5528094
        Canada          Chile          China       Colombia     Costa Rica
    -0.3168924     -8.2422307      2.5360361     -1.4517071      5.1250782
       Denmark        Ecuador        Finland         France        Germany
     5.4002388     -2.4056313     -1.6810857      2.4754718     -0.1806993
        Greece      Guatamala       Honduras        Iceland          India
    -3.1161685     -3.3552838      0.7100245     -6.2105820      0.5086740
       Ireland          Italy          Japan          Korea     Luxembourg
     3.3911306      1.9267549      5.2814855     -6.1069814     -1.6708066
         Malta         Norway    Netherlands    New Zealand      Nicaragua
     2.9749098     -0.8717854      0.4255455      2.2855548      0.6463966
        Panama       Paraguay           Peru    Philippines       Portugal
    -3.2941656     -6.1257589      6.5394410      6.6750084     -0.7684447
  South Africa South Rhodesia          Spain         Sweden    Switzerland
     0.4831656      1.2914342     -0.6711565     -4.2602834      2.4868259
        Turkey        Tunisia United Kingdom  United States      Venezuela
    -2.6656824     -2.8179200     -2.6924128     -1.1115901      3.6325177
        Zambia        Jamaica        Uruguay          Libya       Malaysia
     9.7509138     -3.0185314     -2.2638273     -2.8295257     -2.9708690
```

```
> m=lm(sr˜pop15+pop75+dpi+ddpi,data=LifeCycleSavings)
> summary(m)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2422 -2.6857 -0.2488  2.4280  9.7509

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.5660865  7.3545161   3.884 0.000334 ***
pop15       -0.4611931  0.1446422  -3.189 0.002603 **
pop75      -1.6914977  1.0835989  -1.561 0.125530
dpi         -0.0003369  0.0009311  -0.362 0.719173
ddpi        0.4096949  0.1961971   2.088 0.042471 *


Residual standard error: 3.803 on 45 degrees of freedom
Multiple R-squared:  0.3385, Adjusted R-squared:  0.2797
```

Summary

1. model specification and assumptions:

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbb{I})$$

Summary

1. model specification and assumptions:

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbb{I})$$

2. check for multicollinearity! new entry

Summary

1. model specification and assumptions:

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbb{I})$$

2. check for multicollinearity! new entry
3. estimate the model parameters

Summary

1. model specification and assumptions:

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbb{I})$$

2. check for multicollinearity! new entry
3. estimate the model parameters
4. diagnostics:
   - R^2 or Adjusted R^2
   - t-test
   - test for homoskedasticity

Summary

1. model specification and assumptions:

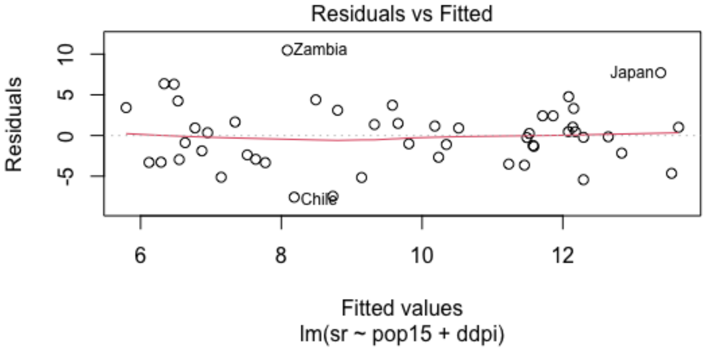
$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 \mathbb{I})$$

2. check for multicollinearity! new entry
3. estimate the model parameters
4. diagnostics:
   - R^2 or Adjusted R^2
   - t-test
   - test for homoskedasticity
5. interpretation

```
> cor(X)
            pop15       pop75        dpi         ddp
pop15  1.00000000 -0.90847871 -0.7561881 -0.04782569
pop75 -0.90847871  1.00000000  0.7869995  0.02532138
dpi   -0.75618810  0.78699951  1.0000000 -0.12948552
ddp   -0.04782569  0.02532138 -0.1294855  1.00000000
```

```
> m=lm(sr˜pop15+ddpi,data=LifeCycleSavings)
> summary(m)
Residuals:
    Min      1Q  Median      3Q     Max
-7.5831 -2.8632  0.0453  2.2273 10.4753

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.59958    2.33439   6.682 2.48e-08 ***
pop15       -0.21638    0.06033  -3.586 0.000796 ***
ddpi         0.44283    0.19240   2.302 0.025837 *
---
Residual standard error: 3.861 on 47 degrees of freedom
Multiple R-squared:  0.2878,^^IAdjusted R-squared:  0.2575
```


Residuals vs Fitted
Residuals
10
5
0
-5
Zambia
Japan
Chile
6
8
10
12
Fitted values
lm(sr ~ pop15 + ddpi)