Applied Statistics in R

Elena Parilina

Master's Program
Game Theory and Operations Research

Saint Petersburg State University

2019

Agenda

1. Linear regression

Linear regression

Linear regression

y: dependent variable,
$x = (x_1, \ldots, x_k)^\intercal$: predictor (explanatory) variables.

The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i is an unobserved random variable.

Linear regression

y: dependent variable,
$x = (x_1, \ldots, x_k)^\intercal$: predictor (explanatory) variables.

The model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, \ldots, n,$$

where ε_i is an unobserved random variable.

Assumptions:

- $E\varepsilon_i = 0, \quad i = 1, \ldots, n,$
- $E(\varepsilon_j \varepsilon_\ell) = 0, \quad j \neq \ell,$
- $E\varepsilon_i^2 = \sigma^2, \quad i = 1, \ldots n,$
- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal \sim N(0, \sigma^2 I_n).$

Linear regression

The model in a matrix form:

$$Y = X\beta + \varepsilon,$$

where $Y = (y_1, \ldots, y_n)^\mathsf{T}$, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\mathsf{T}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\mathsf{T}$,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}.$$

Linear regression

The model in a matrix form:

$$Y = X\beta + \varepsilon,$$

where $Y = (y_1, \ldots, y_n)^\intercal$, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\intercal$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\intercal$,

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{pmatrix}.$$

The problem:

$$\min_{\beta_0, \beta_1, \ldots, \beta_k} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2.$$

MLE estimator: $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Linear regression: $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$.

Properties

Denote $\hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$,

$$S^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} = \frac{(Y-\hat{Y})^\intercal(Y-\hat{Y})}{n-k-1}$$

1. $$\frac{\hat{\beta}_j - \beta_j}{S\sqrt{(X^TX)^{-1}_{j+1,j+1}}} \sim T_{n-k-1}, \quad j = 0, \ldots, k.$$

CI:

$$\left(\hat{\beta}_j - t_{1-\frac{\alpha}{2},n-k-1}S\sqrt{(X^TX)^{-1}_{j+1,j+1}};\right.$$
$$\left.\hat{\beta}_j + t_{1-\frac{\alpha}{2},n-k-1}S\sqrt{(X^TX)^{-1}_{j+1,j+1}}\right),$$

Properties

Denote

$$R^2 = \frac{(\hat{Y} - \bar{y}\mathbf{1})^T(\hat{Y} - \bar{y}\mathbf{1})}{(Y - \bar{y}\mathbf{1})^T(Y - \bar{y}\mathbf{1})} = 1 - \frac{(Y - \hat{Y})^\intercal(Y - \hat{Y})}{(Y - \bar{y}\mathbf{1})^T(Y - \bar{y}\mathbf{1})}.$$

2. $F = \dfrac{R^2}{1 - R^2} \cdot \dfrac{n - k - 1}{k} \sim \mathcal{F}_{k,n-k-1}.$

Using R...

lm

```
 lm(formula, data, subset, weights, na.action, method =
"qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, ...)
```

Using R...

lm

```
lm(formula, data, subset, weights, na.action, method =
"qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, ...)
```

summary

```
summary(object)
```

Using R...

lm

```
lm(formula, data, subset, weights, na.action, method =
"qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, ...)
```

summary

```
summary(object)
```

pairs

```
pairs(formula, data = NULL, ..., subset, na.action =
stats::na.pass)
```

Arguments (lm)

formula: an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

subset: an optional vector specifying a subset of observations to be used in the fitting process.

weights: an optional vector of weights to be used in the fitting process. Should be NULL or a numeric vector. If non-NULL, weighted least squares is used with weights weights (that is, minimizing $sum(w \cdot e^2)$); otherwise ordinary least squares is used.

offset: this can be used to specify an a priori known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector of length equal to the number of cases. One or more offset terms can be included in the formula instead or as well, and if more than one are specified their sum is used.

Details on "lm"

- Models for lm are specified symbolically. A typical model has the form $response \sim terms$ where response is the (numeric) response vector and terms is a series of terms which specifies a linear predictor for response.
- If the formula includes an offset, this is evaluated and subtracted from the response.
- Non-NULL weights can be used to indicate that different observations have different variances (with the values in weights being inversely proportional to the variances); or equivalently, when the elements of weights are positive integers w_i, that each response y_i is the mean of w_i unit-weight observations (including the case that there are w_i observations equal to y_i and the data have been summarized).
- All of weights, subset and offset are evaluated in the same way as variables in formula, that is first in data and then in the environment of formula.

More $lm()$ examples are available e.g., in (datasets) anscombe, attitude,

Using R...

predict

```
 predict(object, newdata, se.fit = FALSE, scale = NULL, df
= Inf, interval = c("none", "confidence", "prediction"),
level = 0.95, type = c("response", "terms"), terms = NULL,
na.action = na.pass, pred.var = res.var/weights, weights =
1, ...)
```

We have a sample:

```
y <- c(132,143,153,162,154,168,137,149,159,128,166)
x1 <- c(52,59,67,73,64,74,54,61,65,46,72)
x2 <- c(173,184,194,211,196,220,188,188,207,167,217)
```

We want to have the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

We use functions `lm` and `summary`:

```
res<-lm(y ~ x1+x2)
summary(res)
```

Results:

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min	1Q	Median	3Q	Max
-3.4640	-1.1949	-0.4078	1.8511	2.6981

Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$	
(Intercept)	30.9941	11.9438	2.595	0.03186	*
x1	0.8614	0.2482	3.470	0.00844	**
x2	0.3349	0.1307	2.563	0.03351	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.318 on 8 degrees of freedom

Multiple R-squared: 0.9768, Adjusted R-squared: 0.9711

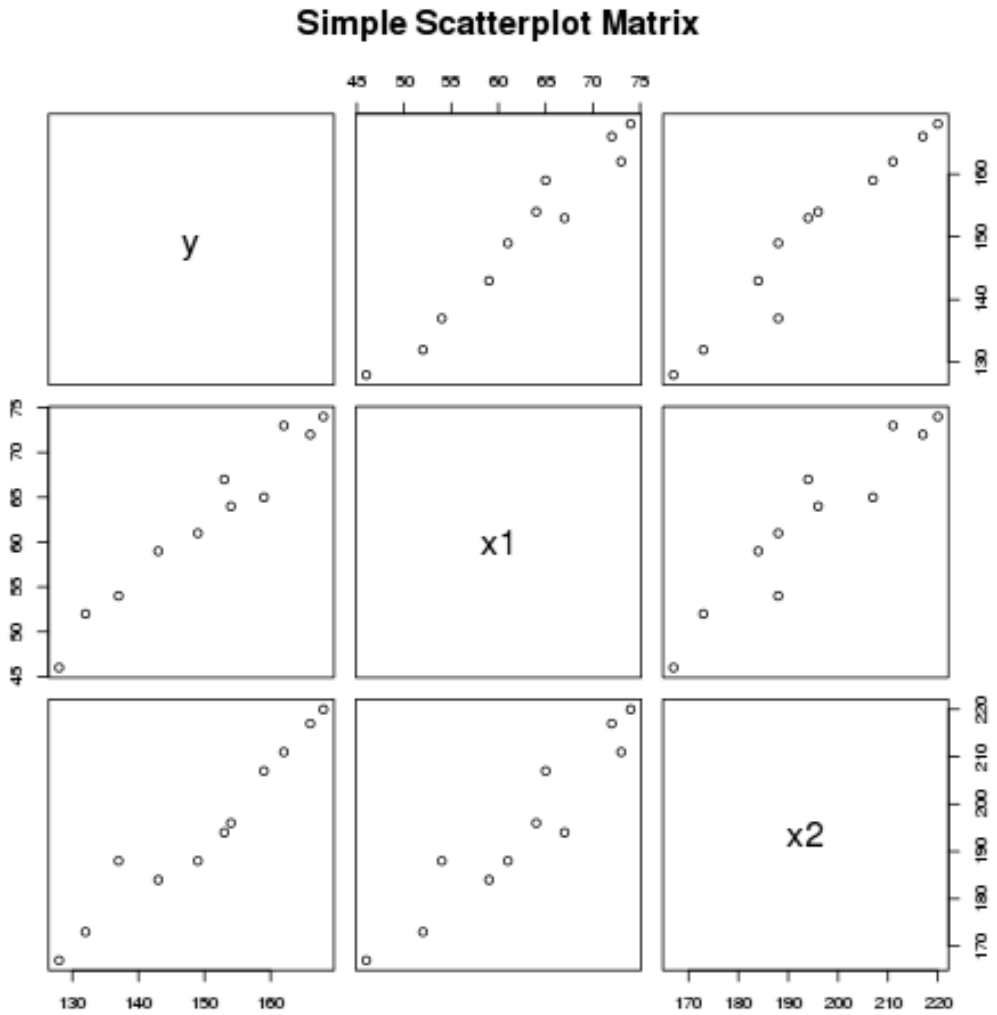
F-statistic: 168.8 on 2 and 8 DF, p-value: 2.874e-07

- We obtain quantiles of levels 0.25, 0.5 and 0.75, estimators β_0, β_1 and β_2, their standard errors $\hat{\beta}_0/t_{\beta_0}$, $\hat{\beta}_1/t_{\beta_1}$, $\hat{\beta}_2/t_{\beta_2}$, values of statistics t_{β_0}, t_{β_1}, t_{β_2} for hypothesis $H_0 : \beta_i = 0$, $i = 0, 1, 2$.
- In column $Pr(>|t|)$ we obtain corresponding p-values. If $p - value > \alpha$ (by default, $\alpha = 0.05$), then $H_0 : \beta_i = 0$ is accepted and coefficient is not significant. Otherwise, null hypothesis is rejected and coefficient is accepted to be significant. In example, all coefficients are significant. Value `Residual standard error` is statistics S.
- `Multiple R-squared` is the value of R^2. Statistics F for hypothesis $H_0 : \beta_1 = \ldots = \beta_k = 0$, is 168.8. And p-value is 2.874e-07, which is less than 0.05. Therefore, the null hypothesis is rejected and we may state that the linear regression model is significant in general.

Use function `pairs`:

```
pairs(y~x1+x2, main="Simple Scatterplot Matrix")
```

Graphs


Simple Scatterplot Matrix
y
x1
x2
45
50
55
60
65
70
75
130
140
150
160
170
180
190
200
210
220

About the model

- `coef(model)` gives the coefficients of linear regression.
- `fitted(model)` gives $\hat{y}$ values.
- `summary(model)` gives the summary of the model.
- `confint(model,"variable")` gives confident interval for "variable".
- `anova(model)` gives "analysis of variances table".

```
 > x<-c(1,2,3,4,5,6)
> y<-c(2,4,5.5,8,10.3,11.7)
> mmodel<-lm(y ~ x)
> summary(mmodel)
```

Result:

```
Call:
lm(formula = y ~ x)
Residuals:
1 2 3 4 5 6
0.07619 0.07905 -0.41810 0.08476 0.38762 -0.20952
Coefficients:
Estimate   Std.      Error t value Pr(>|t|)
(Intercept) -0.07333 0.29001 -0.253 0.813
x       1.99714 0.07447 26.819 1.15e-05 ***
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1
‘ ’ 1
Residual standard error:  0.3115 on 4 degrees of freedom
Multiple R-squared:  0.9945, Adjusted R-squared:  0.9931
F-statistic:  719.2 on 1 and 4 DF, p-value:  1.149e-05
```

> anova(mmodel)

```
 Analysis of Variance Table
Response:  y
Df Sum Sq Mean Sq F value Pr(>F)
x 1 69.800 69.800 719.24 1.149e-05 ***
Residuals 4 0.388 0.097
---
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1
‘ ’ 1
```

Test for heteroscedasticity

Homoscedasticity of the linear regression model means that ε_t and ε_k have the same variance ofr any t and k.
H_0: the variance of the residuals is constant (homoscedasticity).
H_1: the variances of the residuals are different (heteroscedasticity).

"Car" library

```
 library(car)
ncvTest(mmodel)
```

Result

```
 > ncvTest(mmodel)
Non-constant Variance Score Test
Variance formula:     fitted.values
Chisquare = 0.3530488, Df = 1, p = 0.55239
```

Test for heteroscedasticity

"lmtest" library

```
 > library(lmtest)
> bptest(mmodel)
```

Result

```
 studentized Breusch-Pagan test
data:  mmodel
BP = 0.58981, df = 1, p-value = 0.4425
```

Test for autocorrelation

Autocorrelation means that the residuals satisfy the equation:

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t,$$

where $\{\nu_t\}$ is the independent normally $(0, \sigma_\nu^2)$ distributed random variables, and $|\rho| < 1$ is an autocorrelation parameter.
H_0: $\rho = 0$ (autocorrelation is zero).
H_1: $\rho \neq 0$ (autocorrelation is not zero).

"Car" library

```
 > library("car")
> durbinWatsonTest(mmodel)
```

Result

```
 lag Autocorrelation D-W Statistic p-value
1 -0.2854923 2.442941 0.922
Alternative hypothesis:  rho != 0
```

Test for autocorrelation

"lmtest" library

```
 > library("lmtest")
> dwtest(mmodel)
```

Result

```
 Durbin-Watson test
data:  mmodel
DW = 2.4429, p-value = 0.4997
alternative hypothesis:  true autocorrelation is greater
than 0
```

Stepwise selection of the model

Step function

```
step(object, scope, scale = 0, direction = c("both",
"backward", "forward"), trace = 1, keep = NULL, steps =
1000, k = 2, ...)
```

Read info here:
`https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/step`

- object — model of `lm`
- scope — defines the range of models examined in the stepwise search. This should be either a single formula, or a list containing components upper and lower, both formulae.
- scale — used in the definition of the AIC statistic for selecting the models

stepAIC() [MASS package]

The function chooses the best model by AIC. It has an option named direction, which can take the following values: i) "both" (for stepwise regression, both forward and backward selection); "backward" (for backward selection) and "forward" (for forward selection). It return the best final model.

stepAIC()

```
 library(MASS)
# Fit the full model
full.model <- lm(y ~., data = swiss)
# Stepwise regression model
step.model <- stepAIC(full.model, direction = "both",
trace = FALSE)
summary(step.model)
```

Datasets

1. [cigarettes.dat.txt] NAME: Cigarette data for an introduction to multiple regression
TYPE: A sample of 25 brands of cigarettes
SIZE: 25 observations, 5 variables.
DESCRIPTIVE ABSTRACT:
Measurements of weight and tar, nicotine, and carbon monoxide content are given for 25 brands of domestic cigarettes.
SOURCES:
Mendenhall, William, and Sincich, Terry (1992), "Statistics for Engineering and the Sciences" (3rd ed.), New York: Dellen Publishing Co.
VARIABLE DESCRIPTIONS:
Brand name
Tar content (mg)
Nicotine content (mg)
Weight (g)
Carbon monoxide content (mg)

Values are delimited by blanks. There are no missing values.
SPECIAL NOTES:
Observation 3 (Bull Durham) is an outlying point.
STORY BEHIND THE DATA:
The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide content. The United States Surgeon General considers each of these substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke.
2. Systolic Blood Pressure Data [mlr02.xls]
The data (X1, X2, X3) are for each patient.
X1 = systolic blood pressure
X2 = age in years
X3 = weight in pounds

3. [kuiper.xls] NAME: Car Data
TYPE: Multiple Regression
SIZE: 810 observations, 12 variables
DESCRIPTIVE ABSTRACT:
Data collected from Kelly Blue Book for several hundred 2005 used GM cars allows students to develop a multivariate regression model to determine their car value based on a variety of characteristics such as mileage, make, model, engine size, interior style, and cruise control.
SOURCES:
For this data set, a representative sample of over eight hundred, 2005 GM cars were selected, then an algorithm was developed following the 2005 Central Edition of the Kelly Blue Book to estimate retail price.
VARIABLE DESCRIPTIONS:
Price: suggested retail price of the used 2005 GM car in excellent condition. The condition of a car can greatly affect price. All cars in this data set were less than one year old when priced and considered to be in excellent condition.

Mileage: number of miles the car has been driven
Make: manufacturer of the car such as Saturn, Pontiac, and Chevrolet
Model: specific models for each car manufacturer such as Ion, Vibe, Cavalier
Trim (of car): specific type of car model such as SE Sedan 4D, Quad Coupe 2D
Type: body type such as sedan, coupe, etc.
Cylinder: number of cylinders in the engine
Liter: a more specific measure of engine size
Doors: number of doors
Cruise: indicator variable representing whether the car has cruise control ($1 =$ cruise)
Sound: indicator variable representing whether the car has upgraded speakers ($1 =$ upgraded)
Leather: indicator variable representing whether the car has leather seats ($1 =$ leather)