Linear regression

Jose M Sallan
jose.maria.sallan@upc.edu

Quantitative Research Methods

April 17, 2018

Outline

Regression analysis

Linear regression analysis is about examining the relationship between:

- a **dependent (endogenous, response, criterion)** variable y.
- a set of p **independent (exogenous, predictor)** variables x_j.

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \tag{1}$$

Example: mtcars 1

The `mtcars` dataset comprises fuel comsumption in miles per gallon `mpg` and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

```
> head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Example: mtcars 1

Linear model results

Regression of mpg on horsepower hp weight wt and acceleration: qsec

```
> mtcars01 <- lm(mpg ~ hp + wt + qsec, data=mtcars)
> summary(mtcars01)

Call:
lm(formula = mpg ~ hp + wt + qsec, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8591 -1.6418 -0.4636  1.1940  5.6092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.61053    8.41993   3.279  0.00278 **
hp          -0.01782    0.01498  -1.190  0.24418
wt          -4.35880    0.75270  -5.791 3.22e-06 ***
qsec         0.51083    0.43922   1.163  0.25463
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,        Adjusted R-squared:  0.8171
F-statistic: 47.15 on 3 and 28 DF,  p-value: 4.506e-11
```

Understanding regression analysis output

Overall model significance

- **Coefficient of determination** R^2: fraction of variability of y explained by regression model (R^2 closer to one shows good fit).
- **Adjusted** R^2: adjusted by the number of predictors p.
- **F-statistic**: tests whether the regression model explains y better than $\bar{y}$ (small p-value shows good fit).

Understanding regression analysis output

Regression coefficient significance

- Null hypothesis for each variable: **regression coefficient equals zero** (i.e., x_j is unrelated to y).
- Null hypothesis can be discarded if p-value is small enough (p-value: probability of rejecting null hypothesis being true).

.	$p < 0.1$
*	$p < 0.05$
**	$p < 0.01$
***	$p < 0.001$

Example: mtcars 1

Linear model interpretation

```
> mtcars01 <- lm(mpg ~ hp + wt + qsec, data=mtcars)
> summary(mtcars01)

Call:
lm(formula = mpg ~ hp + wt + qsec, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8591 -1.6418 -0.4636  1.1940  5.6092

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 27.61053    8.41993   3.279  0.00278 **
hp          -0.01782    0.01498  -1.190  0.24418
wt          -4.35880    0.75270  -5.791 3.22e-06 ***
qsec         0.51083    0.43922   1.163  0.25463
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.578 on 28 degrees of freedom
Multiple R-squared:  0.8348,        Adjusted R-squared:  0.8171
F-statistic: 47.15 on 3 and 28 DF,  p-value: 4.506e-11
```

Overall significance?
Significant regression coefficients?

Residuals diagnostics

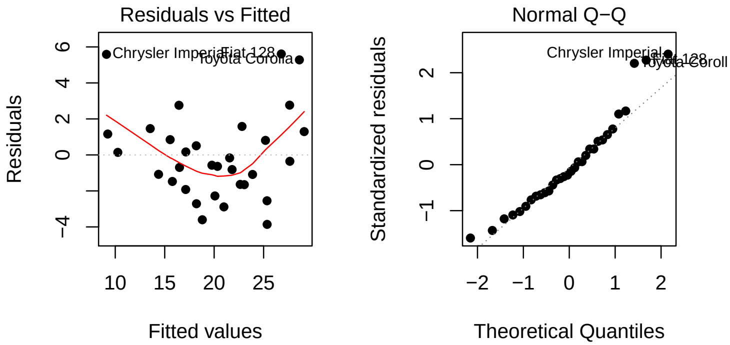
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

- Residuals are the difference between actual and predicted value: $e_i = y_i - \hat{y}_i$.
- Residuals should have mean zero and constant variance across $\hat{y}_i$.
- Residuals should be normally distributed.
- Residuals should be independent of time (if applicable).

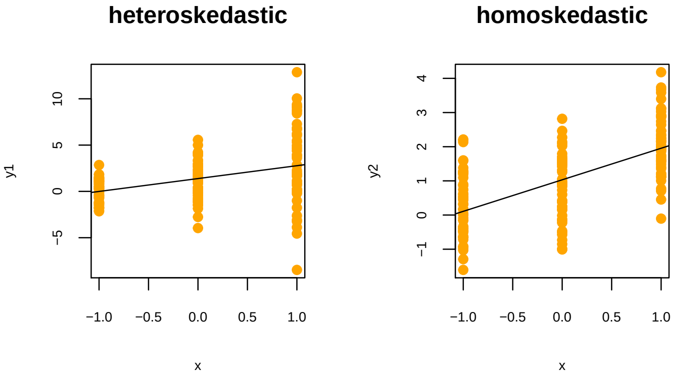
Example: mtcars 1

Residual diagnostics plots

```
> par(mfrow=c(1,2), pty="s")
> plot(mtcars01, which=1, pch=16)
> plot(mtcars01, which=2, pch=16)
```



Example: two regression models



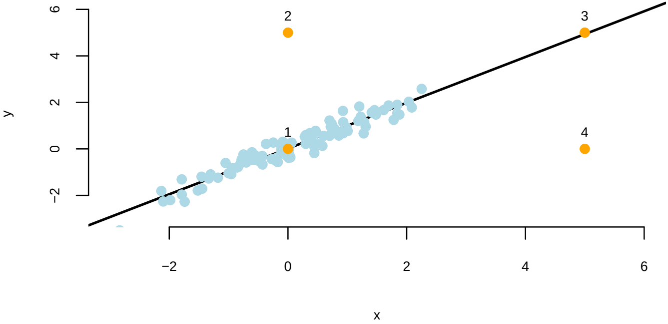
Example: two regression models

Comparing coefficients of model 1 (heteroskedastic) and model 2 (homoskedastic). Real regression coefficient of x is 1.

```
> coef(summary(mod.y1))

             Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 1.385123  0.2333131 5.936757 1.984273e-08
x           1.390449  0.2857490 4.865981 2.885719e-06
```

```
> coef(summary(mod.y2))

             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) 1.0312575 0.07809141 13.205774 8.432673e-27
x           0.9233003 0.09564205  9.653707 2.100486e-17
```

Influential, high-leveraging and outlying points



Point 1 would be not an outlier, the rest of points are
Point 2 would be a low-leverage, low-influence point
Point 3 would be a high-leverage, low-influence point
Point 4 would be a high-leverage, high-influence point

Diagnostics influence

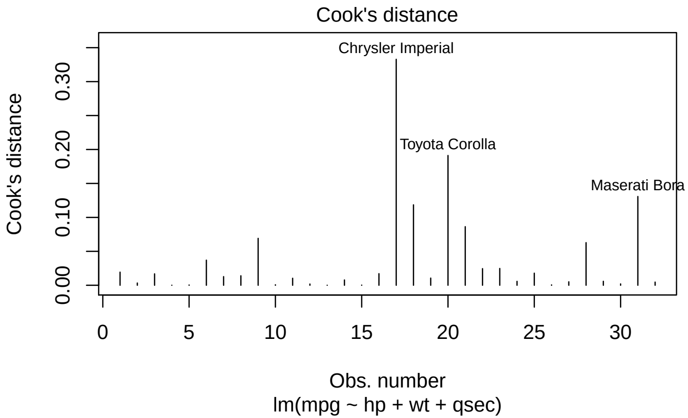
A collection of available influence diagnostics in R can be obtained typing `?influence.measures`

- `cooks.distance` Cook's distance of point i is the standardized variation of predicted values when deleting observation i
- `dfbeta` / `dfbetas` change in unstandardized / standardized regression coefficients when deleting observation i

Example: mtcars 1

Cook's distance plot

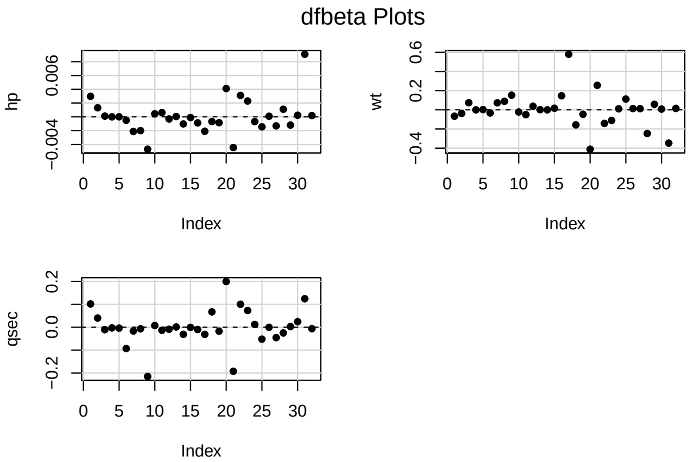
```
> plot(mtcars01, which=4)
```



Example: mtcars 1

Dfbeta plot

```
> library(car)
> dfbetaPlots(mtcars01, pch=16)
```



Multicollinearity

There is **multicollinearity** if some of the predictor variables x_j are highly correlated.

Symptoms of multicollinearity:

- The adjusted coefficient of determination $R^2_{aj.}$ decreases when new variables are added to the model.
- The variance of the coefficient estimators increases when new variables are added to the model.
- A regression coefficient ceases to be significant when new variables are added to the model.

Multicollinearity diagnostics

Multicollinearity (MC) is assessed by regressing each x_j upon the rest of variables, and obtaining coefficients of determination R_j^2.

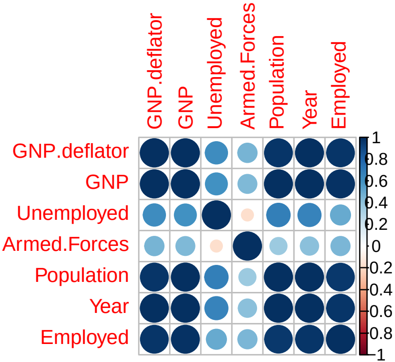
- **Tolerance** of x_j is equal to $1 - R_j^2$ (high if not MC).
- **Variance inflation factor** is the inverse of tolerance (low if not MC).

Example: Longley

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

`longley` is a macroeconomic data set which provides a well-known example for a highly collinear regression.

```
> library(corrplot)
> corrplot(cor(longley), method="circle")
```



Example: Longley

Calculating variance inflation factors

```
> longley01 <- lm(Employed ~ ., data=longley)
> library(car)
> vif(longley01)
```

```
GNP.deflator          GNP    Unemployed Armed.Forces   Population         Year
   135.53244   1788.51348     33.61889      3.58893    399.15102    758.98060
```

```
> longley02 <- lm(Employed ~ GNP + Armed.Forces, data=longley)
> vif(longley02)
```

```
         GNP Armed.Forces
    1.248916     1.248916
```

Example: Longley

Comparing regression coefficients of both models


```
> coef(summary(longley01))
                  Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)  -3.482259e+03 8.904204e+02 -3.9108029 0.0035604037
GNP.deflator  1.506187e-02 8.491493e-02  0.1773760 0.8631408328
GNP          -3.581918e-02 3.349101e-02 -1.0695163 0.3126810611
Unemployed   -2.020230e-02 4.883997e-03 -4.1364274 0.0025350917
Armed.Forces -1.033227e-02 2.142742e-03 -4.8219853 0.0009443668
Population   -5.110411e-02 2.260732e-01 -0.2260511 0.8262117958
Year          1.829151e+00 4.554785e-01  4.0158898 0.0030368033
> coef(summary(longley02))
                Estimate  Std. Error    t value     Pr(>|t|)
(Intercept)  51.683468731 0.804340962 64.2556716 1.164460e-17
GNP           0.034393472 0.001965575 17.4979215 2.036369e-10
Armed.Forces  0.001147955 0.002807338  0.4089122 6.892606e-01
```

Categorical variables

In a regression model, we can be interested in introducing **categorical variables** (including different levels):

- gender
- industry

A categorical variable with k levels can be modelled through $k-1$ **dummy** variables

Sector	d_1	d_2	d_3
Energy (base level)	0	0	0
Pharmacy	1	0	0
IT	0	1	0
Construction	0	0	1

If the categorical variable is coded with text, R generates the dummy variables automatically. Otherwise, these variables must be specified as factors.
Coefficients of dummy variables represent the difference of value of the response between the value defined by the dummy variable and the base level.

Example: mtcars 2


```
> levels(as.factor(mtcars$gear))

[1] "3" "4" "5"

> mtcars02 <- lm(mpg ~ wt + hp + factor(gear), data=mtcars)
> coef(summary(mtcars02))

                  Estimate Std. Error    t value     Pr(>|t|)
(Intercept)    34.87245123 2.58015801 13.5156262 1.558098e-13
wt             -3.23852439 0.87781636 -3.6892960 1.000770e-03
hp             -0.03497069 0.01260201 -2.7750090 9.897557e-03
factor(gear)4   1.26489784 1.34083819  0.9433635 3.538604e-01
factor(gear)5   1.87355541 1.86661986  1.0037156 3.244269e-01
```

In this model, number of gears does not influence fuel consumption

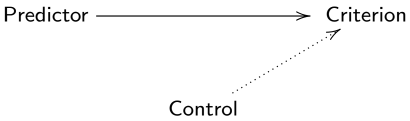
Hierarchical regression

In hierarchical regression independent variables are entered in two steps:

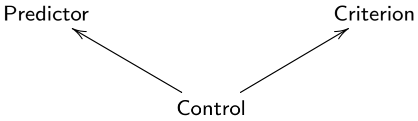
- **Control variables**: sources of variability not directly related with the phenomenon we want to study (e.g. gender, age, etc.).
- **Predictor variables**: variables whose dependence with the **criterion variable** we want to examine.

Hierarchical regression

The aim of hierarchical regression is to prevent spurious correlations with control variables:
Non-spurious relation



Spurious relation



Hierarchical regression

Conditions of hierarchical regression

- The model with all variables has to explain more variability than the model with control variables only (F-test).
- The coefficients of the control variables should not experience significant changes when the predictor variables are introduced.

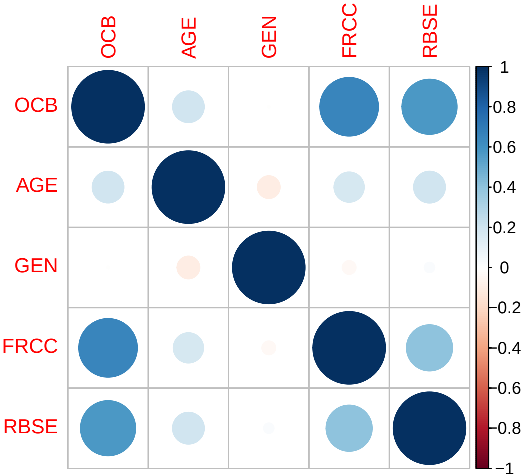
Example: OCB

Dataset with 602 observations from a behavioral study:

- Criterion: Organizational citizenship behavior (OCB).
- Control: AGE and GENder.
- Predictor: Felt responsibility for constructive change (FRCC), role-breadth self-efficacy (RBSE).

Example: OCB

Correlogram of variables



	Dependent variable:	
	OCB	
	(1)	(2)
AGE	0.015***	0.003
	(0.003)	(0.002)
factor(GEN)1	0.028	0.020
	(0.063)	(0.044)
FRCC		0.128***
		(0.008)
RBSE		0.390***
		(0.032)
Constant	3.151***	0.287*
	(0.141)	(0.151)
Observations	602	602
R2	0.036	0.544
Adjusted R2	0.033	0.541
Residual Std. Error	0.772 (df = 599)	0.531 (df = 597)
F Statistic	11.331*** (df = 2; 599)	178.267*** (df = 4; 597)
Note:		*p<0.1; **p<0.05; ***p<0.01

Mediation

A **mediator** is a variable that accounts for the relationship between **predictor** and **criterion**:

$$\begin{tikzcd} \text{Predictor} \arrow[rr, dashed, "c"] \arrow[dr, "a"] & & \text{Criterion} \\ & \text{Mediator} \arrow[ur, "b"] & \end{tikzcd}$$

Mediation

Baron and Kenny criteria: mediation exists when:

- variations in the level of the independent variable account for variations in the presumed mediator (path *a*).
- variations in the mediator significantly account for variations of the dependent variable (path *b*).
- when paths *a* and *b* are controlled, a previously significant relationship between the dependent and independent variable (path *c*) is no longer significant.

Mediation

A simulation of a mediated relationship

```
> set.seed(3333)
> pred <- rnorm(100, 2, 1)
> med <- 3 + 2*pred + rnorm(100, sd=0.3)
> cri <- 2 + med + rnorm(100, sd=0.2)
> bk01 <- lm(med ~ pred)
> bk02 <- lm(cri ~ pred)
> bk03 <- lm(cri ~ pred + med)
```

Mediation

A simulation of a mediated relationship

```
> coef(summary(bk01))

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 3.018608 0.07659320 39.41092 6.544253e-62
pred        2.006344 0.03358632 59.73696 6.803052e-79

> coef(summary(bk02))

            Estimate Std. Error  t value     Pr(>|t|)
(Intercept) 5.010529 0.09142991 54.80186 2.493842e-75
pred        2.006993 0.04009225 50.05938 1.310783e-71

> coef(summary(bk03))

              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 2.10701046 0.22339750  9.4316652 2.264074e-15
pred        0.07714424 0.14597351  0.5284811 5.983726e-01
med         0.96187320 0.07177705 13.4008470 8.487747e-24
```

Moderation

A **moderator** is a variable that affects the direction and/or strength of the relation between **predictor** and **criterion**.

$$\begin{array}{ccc} \text{Predictor} & \longrightarrow & \text{Criterion} \\ & \uparrow & \\ & \text{Moderator} & \end{array}$$

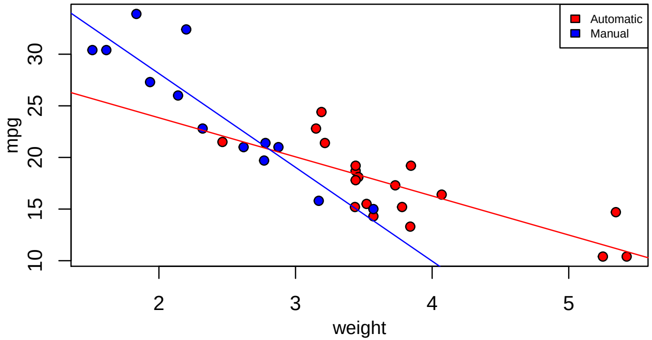
The moderation relationship exists when the coefficient of **interaction term** (product of predictor and moderator) is significant.

Example: mtcars 3

Does the relationship between fuel consumption mpg and weight wt depend on the type of transmission am (0 automatic, 1 manual)?

```
> summary(lm(mpg ~ am*wt, mtcars))

Call:
lm(formula = mpg ~ am * wt, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6004 -1.5446 -0.5325  0.9012  6.0909

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.4161     3.0201  10.402 4.00e-11 ***
am           14.8784     4.2640   3.489  0.00162 **
wt           -3.7859     0.7856  -4.819 4.55e-05 ***
am:wt        -5.2984     1.4447  -3.667  0.00102 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.591 on 28 degrees of freedom
Multiple R-squared:  0.833,       Adjusted R-squared:  0.8151
F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

Example: mtcars 3



Miles per gallon decrease more slowly as weigth increases when automatic gear is used.