Regressione Lineare e Anova

Progetto di Modelli e Metodi dell'Inferenza Statistica

Giacomo Kirn, Francesco Ligorio, Tommaso Parma, Alessandro Papi

Politecnico di Milano

Giugno 2024

Table of Contents

Scelta del dataset: QS World University Rankings 2025

Fonte: QS world university ranking 2025
[https://www.topuniversities.com/world-university-rankings].
Dataset preso da:
[https://www.kaggle.com/datasets/darrylljk
/worlds-best-universities-qs-rankings-2025/data]

Covariate

- 2025 Rank (discreta)
- Institution Name (categorica)
- Location (categorica)
- Size (categorica)
- Academic Reputation (continua)
- Faculty Student (continua)
- Citations per Faculty (continua)
- International Students (continua)
- Employment Outcomes (continua)
- Sustainability (continua)
- QS overall score (continua)

con 595 osservazioni.

Questi indici sono stati usati da Quacquarelli Symonds (QS), una società di consulenza nel settore dell'istruzione, per stilare una classifica delle migliori università a livello mondiale.

Vogliamo interpretare come viene influenzato l'overall delle migliori università mondiali e vogliamo fare previsione del **QS overall score** del *Politecnico di Milano*. Inoltre vogliamo capire se le università Americane ed Europee sono migliori di quelle italiane.

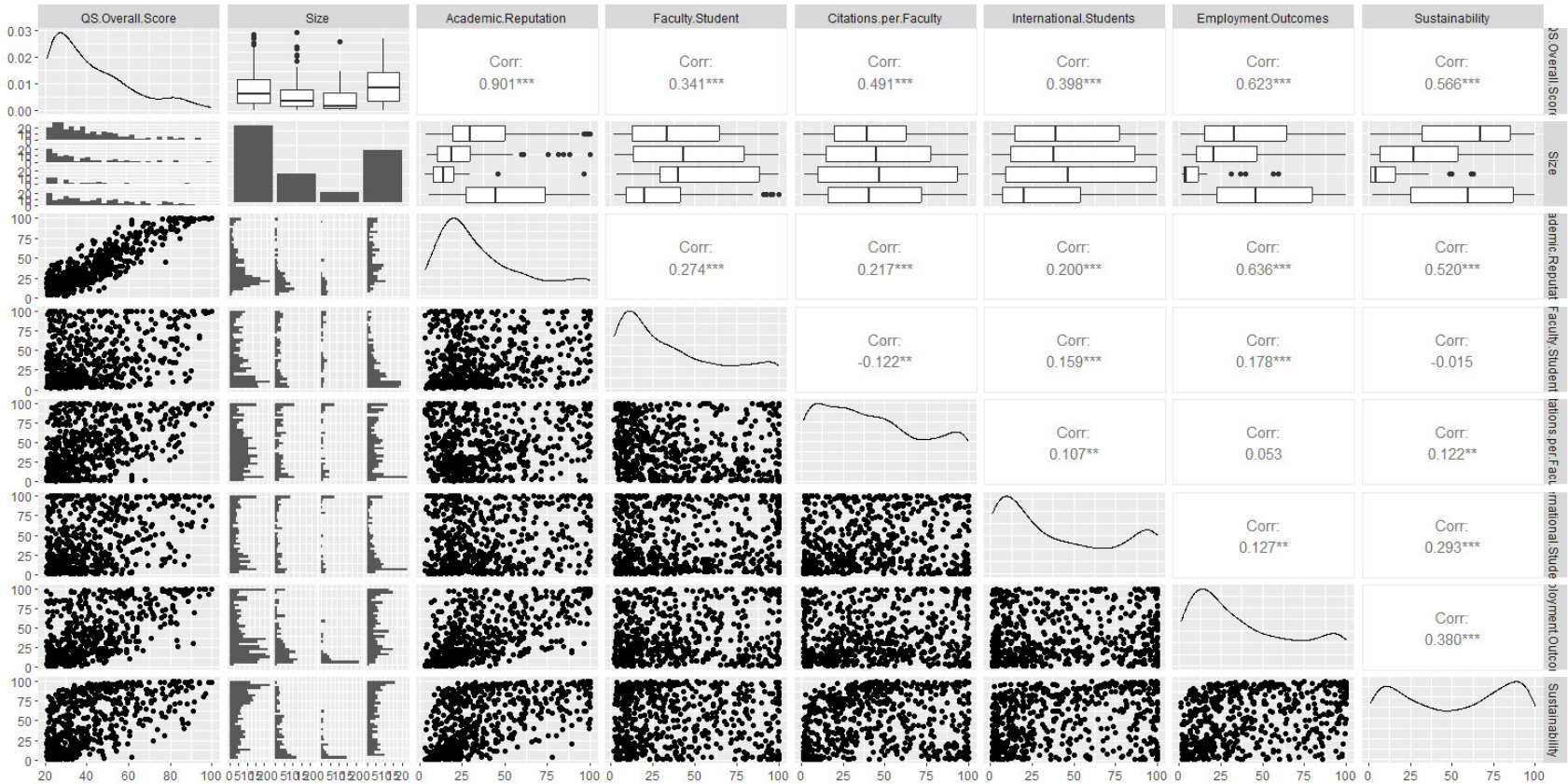
Overview dei dati

	2025.Rank	Institution.Name	Location	Size	Academic.Reputation	Faculty.Student
1	1	Massachusetts Institute of Technology (MIT)	US	M	100.0	100.0
2	2	Imperial College London	UK	L	98.5	98.2
3	3	University of Oxford	UK	L	100.0	100.0
4	4	Harvard University	US	L	100.0	96.3
5	5	University of Cambridge	UK	L	100.0	100.0
6	6	Stanford University	US	L	100.0	100.0

	Citations.per.Faculty	International.Students	Employment.Outcomes	Sustainability	QS.Overall.Score
1	100.0	86.8	100.0	99.0	100.0
2	93.9	99.6	93.4	99.7	98.5
3	84.8	97.7	100.0	85.0	96.9
4	100.0	69.0	100.0	84.4	96.8
5	84.6	94.8	100.0	84.8	96.7
6	99.0	60.8	100.0	81.2	96.1

Ci sono degli NA, per cui tolgiamo le righe in cui sono presenti, arrivando a 587 ossrervazioni.

Primo *ggpairs* (togliendo *Institution Name*, *Rank 2025* e *Location*, per non appesantire troppo il grafico)



Generiamo il primo modello lineare, come risposta `QS.Overall.Score` e come covariate le stesse del *ggpairs* (dataset con l'80% dei dati = 469 osservazioni).

```
Call:
lm(formula = QS.Overall.Score ~ Size + Academic.Reputation +
    Faculty.Student + Citations.per.Faculty + International.Students +
    Employment.Outcomes + Sustainability, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3291 -1.8006 -0.1858  1.7555  8.2446

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.843376   0.369787   7.689 9.09e-14 ***
SizeM                   0.547339   0.362293   1.511   0.1315
SizeS                  -0.117019   0.562815  -0.208   0.8354
SizeXL                 -0.722049   0.308647  -2.339   0.0197 *
Academic.Reputation     0.439538   0.007459  58.927  < 2e-16 ***
Faculty.Student         0.100726   0.004642  21.698  < 2e-16 ***
Citations.per.Faculty   0.200287   0.004173  47.995  < 2e-16 ***
International.Students  0.086636   0.003841  22.553  < 2e-16 ***
Employment.Outcomes     0.074830   0.005251  14.251  < 2e-16 ***
Sustainability          0.064085   0.004790  13.380  < 2e-16 ***
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.595 on 459 degrees of freedom
Multiple R-squared:  0.9814,Adjusted R-squared:  0.9811
F-statistic:  2695 on 9 and 459 DF,  p-value: < 2.2e-16
```

R^2_{adj} iniziale ottimo: 0.9811.
Tutte le covariate molto significative tranne `Size`.
p-value dell'F-test 2.2e-16, c'è evidenza per dire che almeno una covariata sia significativa. Notiamo una leggera asimmetria nei residui.

Togliamo Size dal modello

```
Call:
lm(formula = QS.Overall.Score ~ Academic.Reputation + Faculty.Student +
    Citations.per.Faculty + International.Students + Employment.Outcomes +
    Sustainability, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3805 -1.8357 -0.1695  1.7910  8.4921

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.633709   0.329181   8.001 1.01e-14 ***
Academic.Reputation   0.431138   0.006935  62.168  < 2e-16 ***
Faculty.Student       0.105311   0.004298  24.505  < 2e-16 ***
Citations.per.Faculty 0.202486   0.004086  49.551  < 2e-16 ***
International.Students 0.088763   0.003731  23.791  < 2e-16 ***
Employment.Outcomes   0.075067   0.005250  14.298  < 2e-16 ***
Sustainability        0.064529   0.004570  14.121  < 2e-16 ***
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.614 on 462 degrees of freedom
Multiple R-squared:  0.981,Adjusted R-squared:  0.9808
F-statistic:  3981 on 6 and 462 DF,  p-value: < 2.2e-16
```

Come ci si aspettava l'R^2_{adj} rimane praticamente invariato.
Escludiamo quindi questa covariata.

Cross-Validation

Vediamo ora se stiamo overfittando i dati. L'MSE del test set è `6.811108`, mentre l'MSE sul training set è `6.732053`.
Utilizziamo dunque la **Cross-Validation** con $K = 5$ *folds*.
L'errore di Cross-Validation è `6.888628`, come ci aspettavamo è maggiore del MSE sul training set.

Semplificazione del Modello

Cerchiamo ora di rendere il modello più semplice per una migliore interpretazione e di levare covariate altamente correlate tra loro (ggpairs). Applichiamo un metodo **Stepwise** direzione Both.

```
Start:  AIC=908.33
QS.Overall.Score ~ Academic.Reputation + Faculty.Student + Citations.per.Faculty +
    International.Students + Employment.Outcomes + Sustainability

                         Df Sum of Sq     RSS     AIC
<none>                                 3157.3  908.33
- Sustainability          1    1362.8  4520.1 1074.61
- Employment.Outcomes     1    1397.1  4554.5 1078.16
- International.Students  1    3868.2  7025.5 1281.44
- Faculty.Student         1    4103.7  7261.0 1296.91
- Citations.per.Faculty   1   16779.6 19936.9 1770.62
- Academic.Reputation     1   26412.9 29570.2 1955.50

Call:
lm(formula = QS.Overall.Score ~ Academic.Reputation + Faculty.Student +
    Citations.per.Faculty + International.Students + Employment.Outcomes +
    Sustainability, data = data_train)

Coefficients:
           (Intercept)     Academic.Reputation         Faculty.Student   Citations.per.Faculty  International.Students
               2.63371                 0.43114                 0.10531                 0.20249                 0.08876
   Employment.Outcomes          Sustainability
               0.07507                 0.06453
```

Notiamo che dovremmo prendere ancora tutte e 6 le covariate.

Proviamo allora a massimizzare l'R^2_{adj} tramite una selezione delle covariate.

I 5 migliori modelli sono:

1,2,3,4,5,6	1,2,3,4,5	1,2,3,4,6	1,2,3,4	1,2,3,5,6
0.981	0.973	0.972	0.962	0.957

Scegliamo allora il 4° modello.

Modello finale

```
  Call:
lm(formula = QS.Overall.Score ~ Academic.Reputation + Faculty.Student +
    Citations.per.Faculty + International.Students, data = data_train)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1212 -2.6972 -0.0187  2.4106 12.7287

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.439329   0.417135   13.04   <2e-16 ***
Academic.Reputation    0.533100   0.006883   77.45   <2e-16 ***
Faculty.Student        0.092938   0.005875   15.82   <2e-16 ***
Citations.per.Faculty  0.192838   0.005673   33.99   <2e-16 ***
International.Students 0.101901   0.005081   20.05   <2e-16 ***
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.655 on 464 degrees of freedom
Multiple R-squared:  0.9627,Adjusted R-squared:  0.9624
F-statistic:  2998 on 4 and 464 DF,  p-value: < 2.2e-16
```

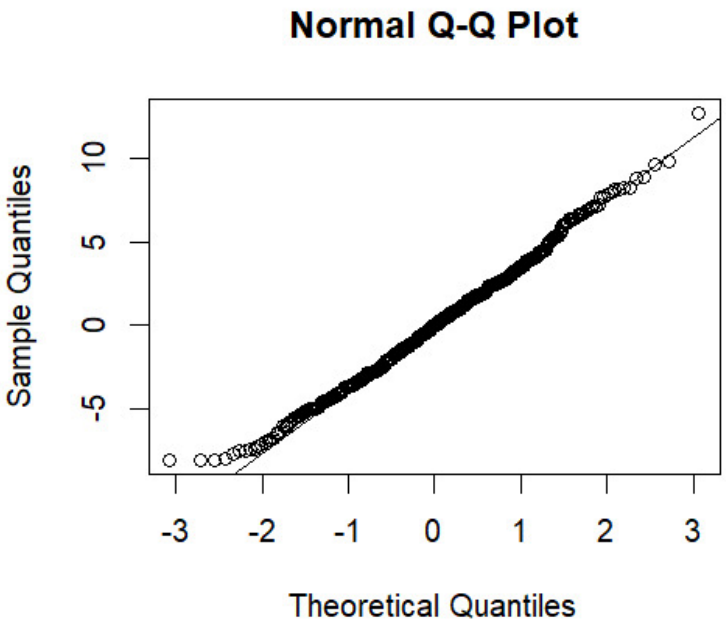
Correlazione tra covariate

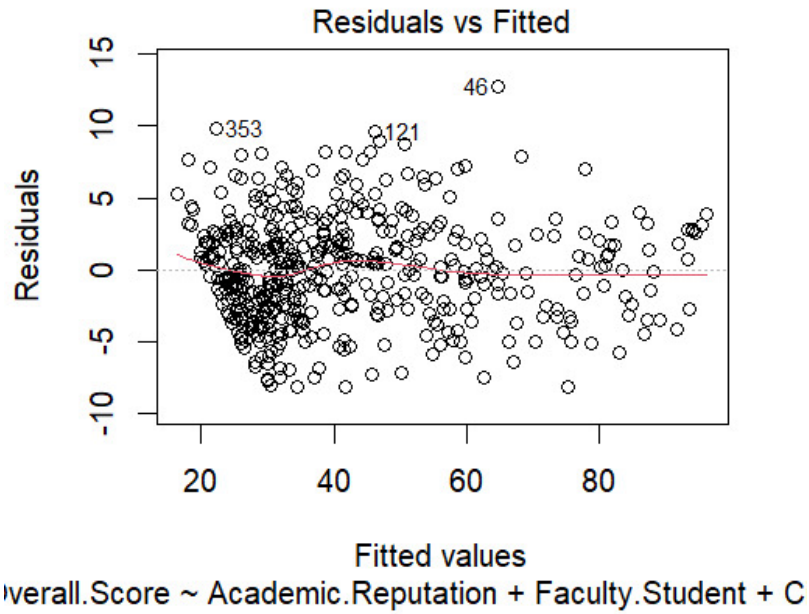
VIF:

Academic.Reputation	Faculty.Student	Citations.per.Faculty	International.Students
1.180027	1.157112	1.080647	1.080132

Controlliamo le ipotesi di validità

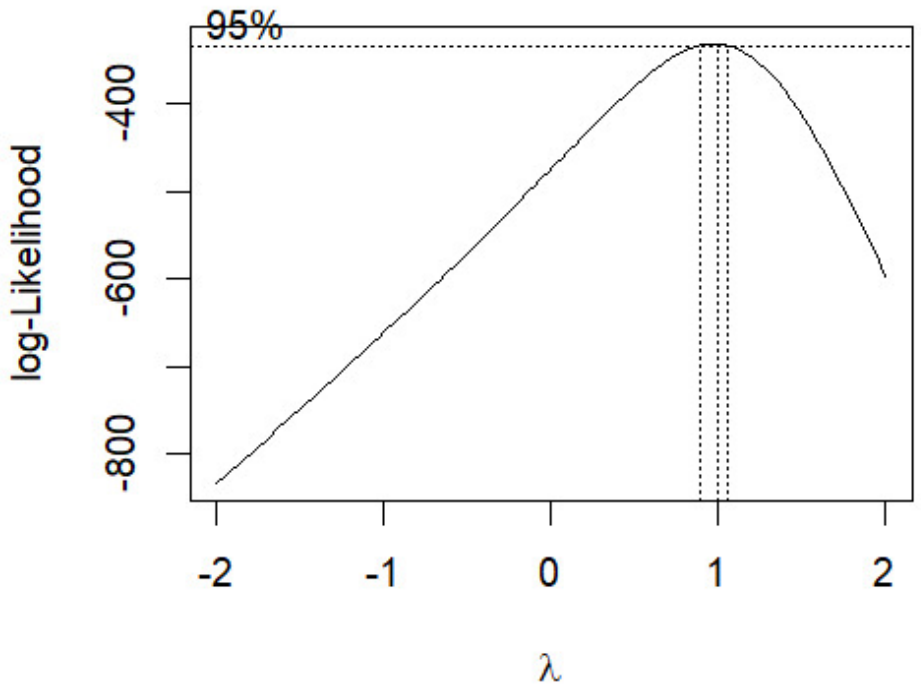
Q-Q plot ottimo e Shapiro Test con p-value: `0.07681`.
Omoschedasticità buona.





Procediamo con una pulizia del dataset individuando punti leva e residui molto grandi. Prima di tutto facciamo un Box Cox.

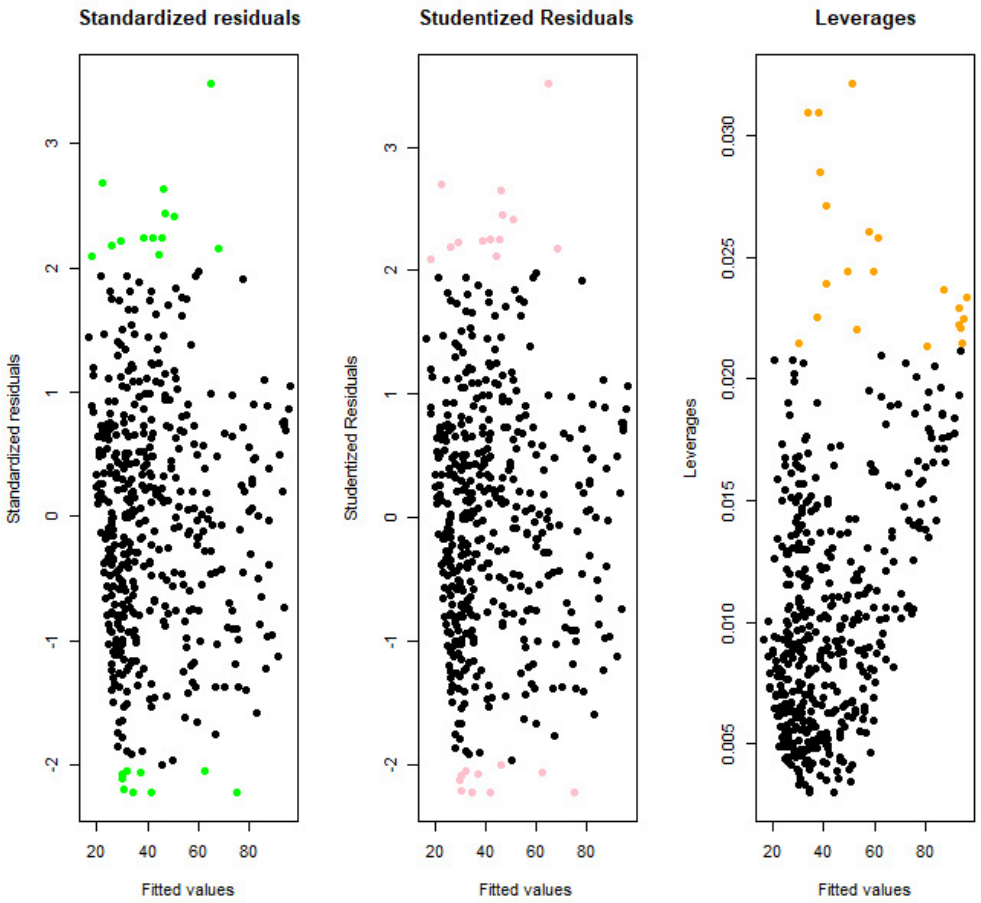
Box Cox



$\lambda_{opt} = 0.989899$

Punti influenti

Punti leva ($lev < 2p/n$), residui standardizzati e studentizzati.



Dopo aver tolto i leverage:

```
        Call:
lm(formula = QS.Overall.Score ~ Academic.Reputation + Faculty.Student +
    Citations.per.Faculty + International.Students, data = data_train,
    subset = (lev < 2 * p/n))

Residuals:
    Min      1Q  Median      3Q     Max
-8.0847 -2.6753  0.0005  2.3802 12.5575

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.553871   0.442143   12.56   <2e-16 ***
Academic.Reputation   0.525918   0.007542   69.73   <2e-16 ***
Faculty.Student       0.095134   0.006520   14.59   <2e-16 ***
Citations.per.Faculty 0.194918   0.006192   31.48   <2e-16 ***
International.Students 0.101780   0.005294   19.23   <2e-16 ***
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.67 on 443 degrees of freedom
Multiple R-squared:  0.9582,Adjusted R-squared:  0.9578
F-statistic:  2539 on 4 and 443 DF,  p-value: < 2.2e-16
```

Dopo aver tolto gli studentizzati:

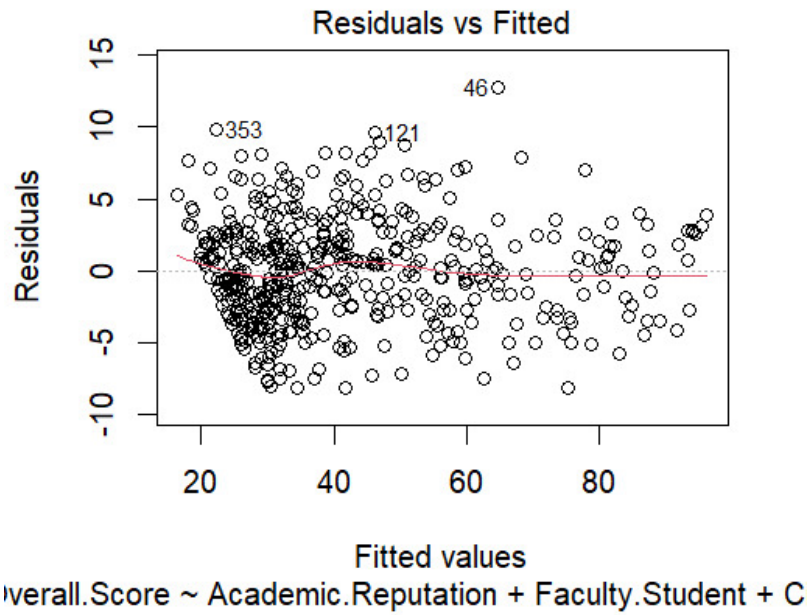
```
      Call:
lm(formula = QS.Overall.Score ~ Academic.Reputation + Faculty.Student +
    Citations.per.Faculty + International.Students, data = data_train,
    subset = (abs(stud) < 2))

Residuals:
    Min      1Q  Median      3Q     Max
-7.2952 -2.4373  0.0352  2.4015  7.3446

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.492925   0.377197   14.56   <2e-16 ***
Academic.Reputation   0.536268   0.006182   86.75   <2e-16 ***
Faculty.Student       0.092893   0.005317   17.47   <2e-16 ***
Citations.per.Faculty 0.187612   0.005121   36.63   <2e-16 ***
International.Students 0.101405   0.004578   22.15   <2e-16 ***
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.217 on 441 degrees of freedom
Multiple R-squared:  0.9714,Adjusted R-squared:  0.9712
F-statistic:  3750 on 4 and 441 DF,  p-value: < 2.2e-16
```

Dopo aver tolto i leverage e gli studentizzati:

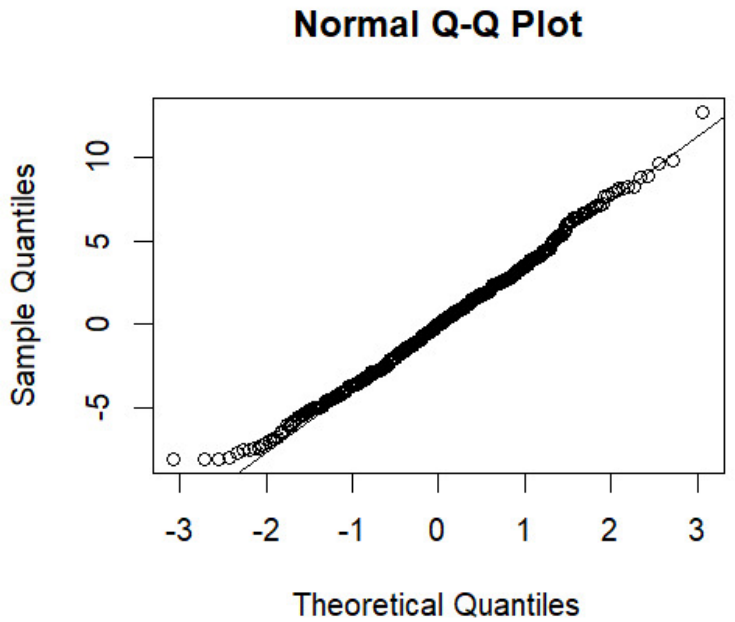
```
        Call:
lm(formula = QS.Overall.Score ~ Academic.Reputation + Faculty.Student +
    Citations.per.Faculty + International.Students, data = data_train,
    subset = (abs(stud) < 2 | lev < 2 * p/n))

Residuals:
    Min      1Q  Median      3Q     Max
-8.1212 -2.6972 -0.0187  2.4106 12.7287

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            5.439329   0.417135   13.04   <2e-16 ***
Academic.Reputation    0.533100   0.006883   77.45   <2e-16 ***
Faculty.Student        0.092938   0.005875   15.82   <2e-16 ***
Citations.per.Faculty  0.192838   0.005673   33.99   <2e-16 ***
International.Students 0.101901   0.005081   20.05   <2e-16 ***
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.655 on 464 degrees of freedom
Multiple R-squared:  0.9627,Adjusted R-squared:  0.9624
F-statistic:  2998 on 4 and 464 DF,  p-value: < 2.2e-16
```

-	post lev.	post stud.	post both
AIC	2443.243	2314.942	2553.622
R^2_{adj}	0.9578	0.9712	0.9624
p-val. ST	0.04692	0.006266	0.07681

Scegliamo il modello trovato senza i leverage e i residui studentizzati (post both).



verall.Score ~ Academic.Reputation + Faculty.Student + C



Interpretazione

$$\texttt{QS.Overall.Score} = 5.439329 + 0.533100 * \texttt{Academic.Reputation} + 0.092938 * \texttt{Faculty.Student} + 0.192838 * \texttt{Citation.Per.Faculty} + 0.101901 * \texttt{International.Student}$$

Previsione

	Institution.Name	Academic.Reputation	Faculty.Student	Citations.per.Faculty	International.Students	QS.Overall.Score
111	Politecnico di Milano	70.8	5.8	40.2	56.8	58.2

- Previsione puntuale: 57.2619
- Intervallo di previsione: $[50.03878, 64.48501]$
- Intervallo di confidenza: $[56.49222, 58.03158]$

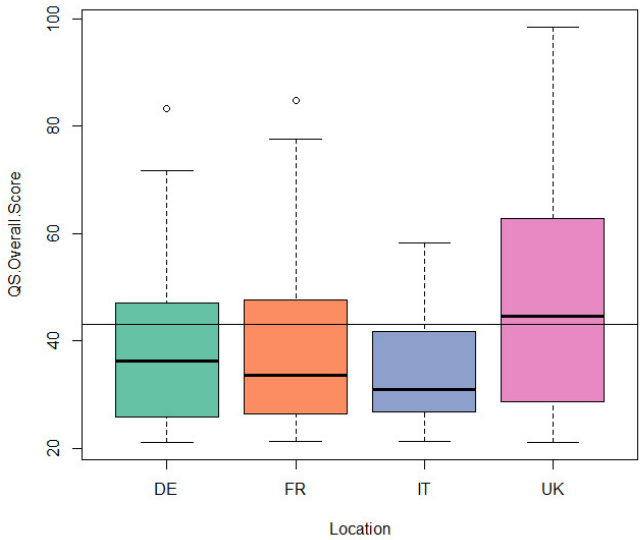
Gli intervalli sono di livello 95%.

Procediamo ora con l'ANOVA, siamo interessati a capire se le università Europee sono sullo stesso livello tra di loro **(A)** e se le università Americane sono migliori di quelle Italiane **(B)**.

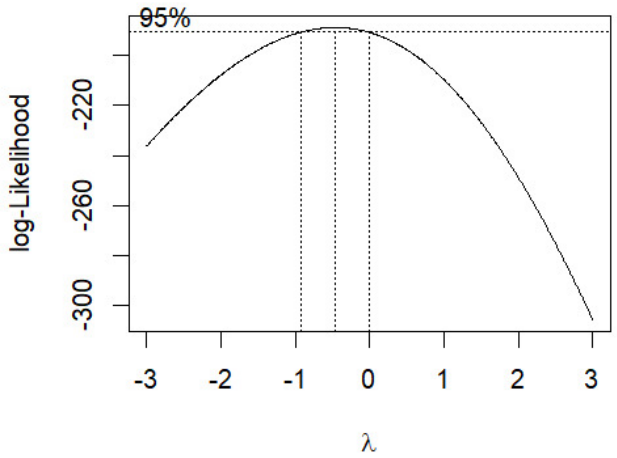
A

Dividiamo le università in base alla `Location`, quelle Europee sono: `UK, FR, DE, IT`.
Lo Shapiro test ci porta tuttavia a rifiutare la normalità nei gruppi.

```
                DE                FR                IT                UK
0.0066716080 0.0074668074 0.0179989071 0.0009852993
```



Procediamo con la trasformazione Box Cox



Otteniamo $\lambda_{opt} = -0.47$
Dallo Shapiro Test non rifiutiamo più la normalità dei gruppi.

```
        DE         FR         IT         UK
0.13134455 0.21398589 0.37093015 0.01169522
```

Il Levene e il Bartlett test confermano la stessa variabilità nei singoli gruppi.

```
        Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  1.5982 0.1934
      121
```

```
        Bartlett test of homogeneity of variances

data:  (dataset_anova$QS.Overall.Score^best_lambda - 1)/best_lambda and dataset_anova$Location
Bartlett's K-squared = 2.5263, df = 3, p-value = 0.4706
```

Genero il modello

```
Call:
lm(formula = (QS.Overall.Score^best_lambda - 1)/best_lambda ~
    Location, data = dataset_anova)

Residuals:
      Min        1Q    Median        3Q       Max
-0.139567 -0.064283 -0.002314  0.056338  0.131952

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.7311392  0.0125875 137.528   <2e-16 ***
LocationFR   0.0004509  0.0218023   0.021   0.9835
LocationIT  -0.0197814  0.0213947  -0.925   0.3570
LocationUK   0.0273904  0.0159576   1.716   0.0886 .
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0734 on 121 degrees of freedom
Multiple R-squared:  0.05483,Adjusted R-squared:  0.0314
F-statistic:  2.34 on 3 and 121 DF,  p-value: 0.07677
```

Procediamo con l'ANOVA

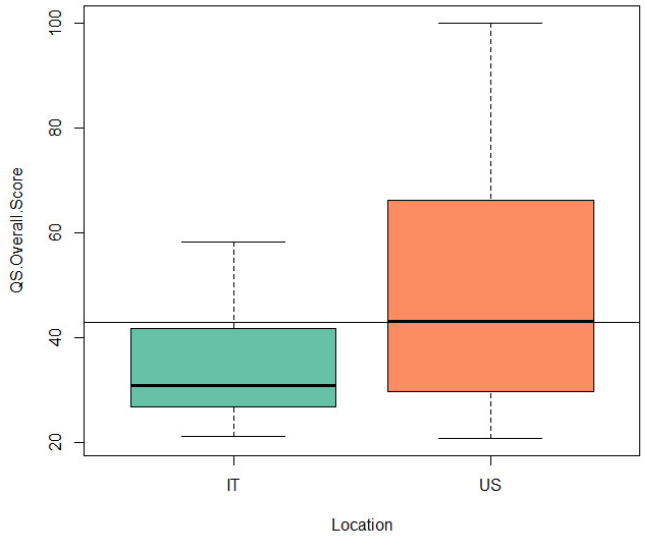
```
        Analysis of Variance Table

Response: (QS.Overall.Score^best_lambda - 1)/best_lambda
           Df  Sum Sq   Mean Sq F value  Pr(>F)
Location    3 0.03782 0.0126058    2.34 0.07677 .
Residuals 121 0.65185 0.0053872
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Il p-value risulta 0.07677, quindi non ho evidenze per dire che in Europa ci siano stati con università migliori di altri.

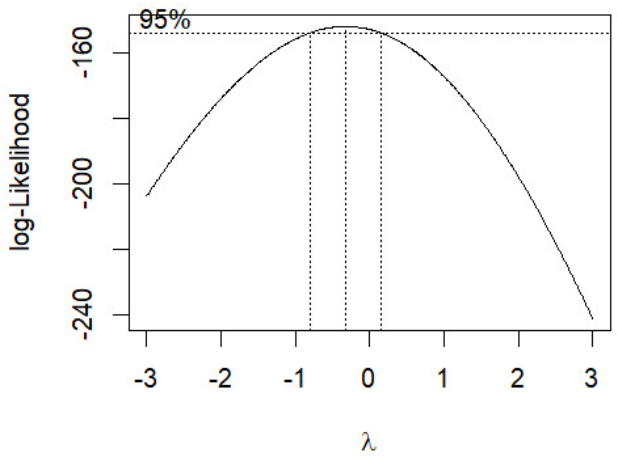
B

Dividiamo le università in base alla `Location`: `IT`, `US`.
Lo Shapiro test ci porta tuttavia a rifiutare la normalità nei gruppi.

```
          IT           US
1.799891e-02 2.057953e-05
```



Procediamo con la trasformazione Box Cox



Otteniamo $\lambda_{opt} = -0.33$
Dallo Shapiro Test non rifiutiamo più la normalità dei gruppi.

```
        IT          US
0.31156126 0.00213126
```

Il Levene e il Bartlett test confermano la stessa variabilità nei singoli gruppi.

```
        Levene's Test for Homogeneity of Variance (center = median)
       Df F value  Pr(>F)
group   1  6.4951 0.01234 *
      100
```

```
        Bartlett test of homogeneity of variances

data:  (dataset_anova_2$QS.Overall.Score^best_lambda - 1)/best_lambda and dataset_anova_2$Location
Bartlett's K-squared = 2.9646, df = 1, p-value = 0.0851
```

Genero il modello

```
Call:
lm(formula = QS.Overall.Score ~ Location, data = dataset_anova_2)

Residuals:
    Min      1Q  Median      3Q     Max
-28.090 -16.590  -4.394  15.060  51.110

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.394      5.057   6.801 7.71e-10 ***
LocationUS    14.496      5.573   2.601   0.0107 *
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.46 on 100 degrees of freedom
Multiple R-squared:  0.06337,Adjusted R-squared:  0.054
F-statistic: 6.766 on 1 and 100 DF,  p-value: 0.0107
```

Procediamo con l'ANOVA

```
        Analysis of Variance Table

Response: QS.Overall.Score
           Df Sum Sq Mean Sq F value Pr(>F)
Location    1   3115 3114.94  6.7658 0.0107 *
Residuals 100  46039  460.39
---
Signif. codes:  0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Il p-value risulta 0.0107, quindi fino al 2% ho evidenze ad affermare che le università Americane siano migliori di quelle Italiane.