Practical session - Modèles de régression linéaire

Piseth KHENG, Borachhun YOU

19 September 2022

IV. Application: GAFAM or BATX data set

The data set[1] below shows the number of monthly active users (MAU) on Facebook from 2008 to 2021 in millions. The numbers were taken from Q4 of each year except for the year 2008, whose data is only available in Q3.

```
tab <- read.table("fb_mau.txt", header=TRUE, sep=",")
tab
```

```
##    year  mau
## 1  2008  100
## 2  2009  360
## 3  2010  608
## 4  2011  845
## 5  2012 1056
## 6  2013 1228
## 7  2014 1393
## 8  2015 1591
## 9  2016 1860
## 10 2017 2129
## 11 2018 2320
## 12 2019 2498
## 13 2020 2797
## 14 2021 2912
```

The dimension of the data set:

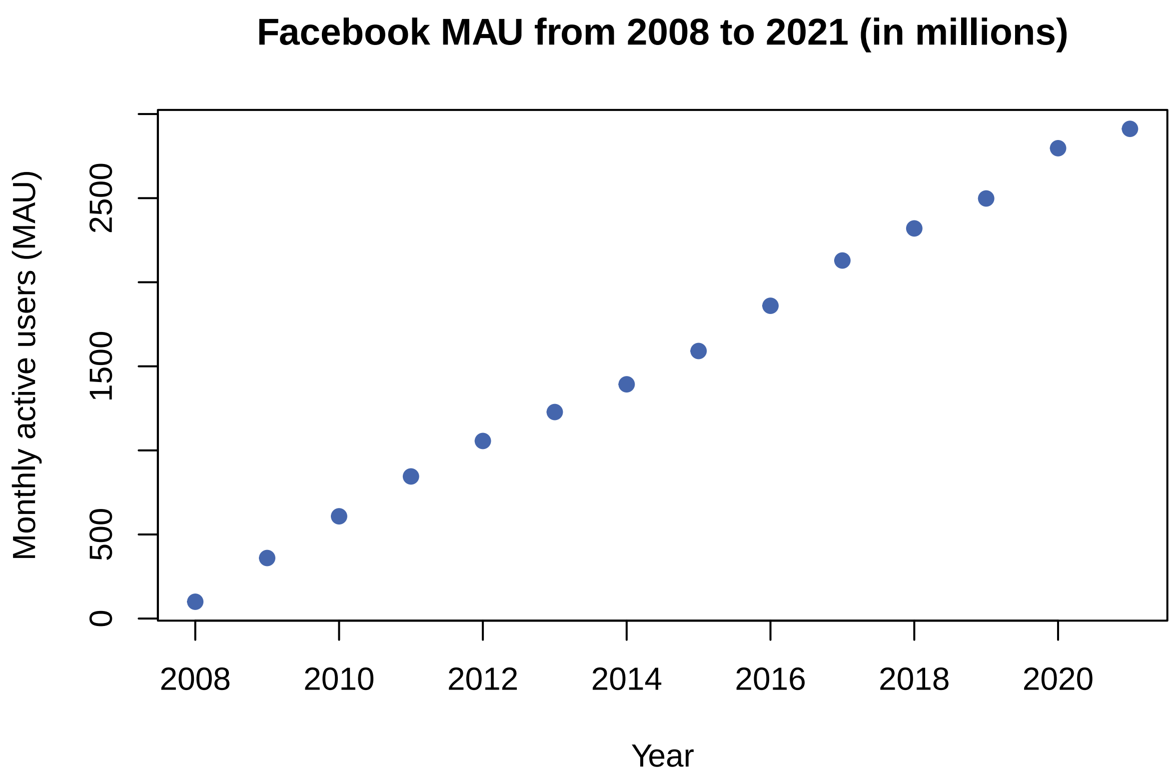
```
dim(tab)
```

```
## [1] 14  2
```

We then try visualizing the data in order to see if there is an apparent linear relationship between the year and the number of users.

```
plot(tab, xlab="Year", ylab="Monthly active users (MAU)",
     main="Facebook MAU from 2008 to 2021 (in millions)",
     pch=19, col=rgb(0.27,0.4,0.68))
```

[1] Source: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/



Based on the graph above, we can see that the relationship is fairly linear. Therefore, we can use a linear model to represent the relationship.

```r
modreg = lm(mau ~ year, data=tab)
summary(modreg)
```

```
##
## Call:
## lm(formula = mau ~ year, data = tab)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.651 -35.664  -0.732  37.167  60.701
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.330e+05  5.764e+03  -75.13   <2e-16 ***
## year         2.157e+02  2.861e+00   75.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.15 on 12 degrees of freedom
## Multiple R-squared:  0.9979, Adjusted R-squared:  0.9977
## F-statistic:  5685 on 1 and 12 DF,  p-value: < 2.2e-16
```

According to the summary of the model, the estimated intercept equals -4.330×10^5 and the estimated coefficient of the year variable equals 2.157×10^2. The model can be written in the form:

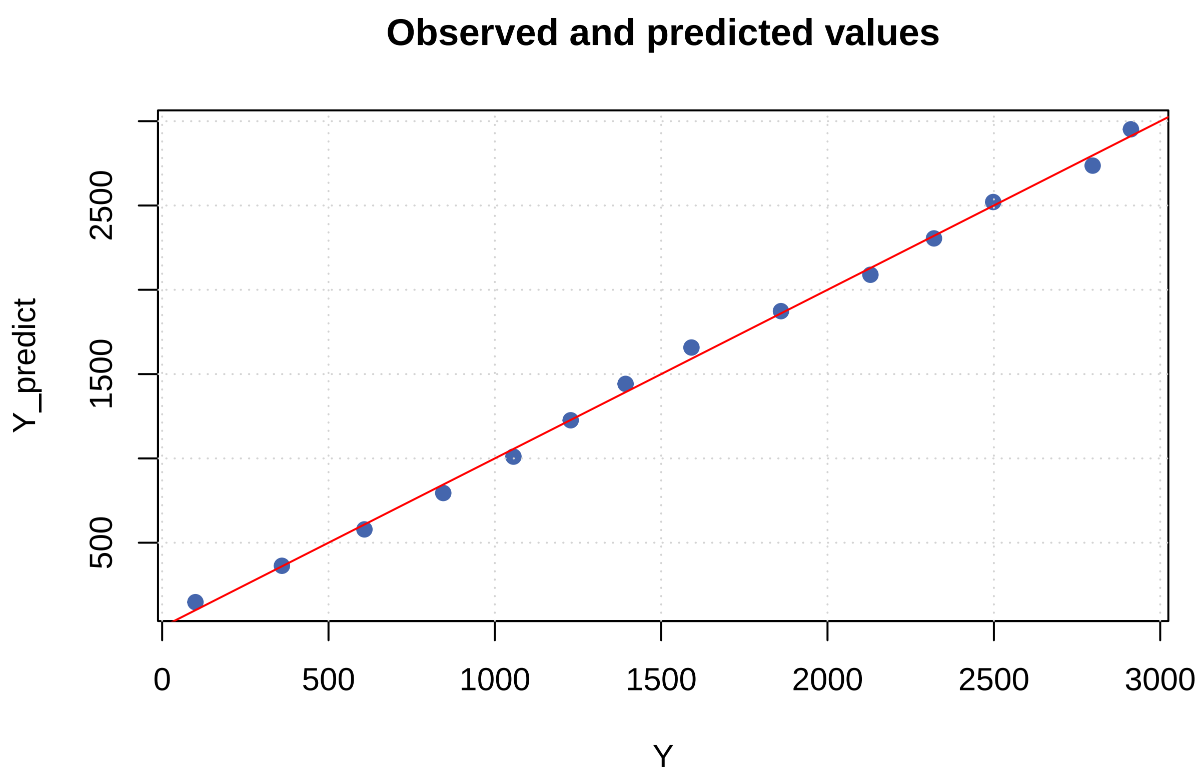
$$\hat{y} = (-4.330 \times 10^5) + (2.157 \times 10^2)x + \hat{\epsilon}$$

where x is the year variable, $\hat{y}$ is the prediction of the MAU and $\hat{\epsilon}$ is the residual.

As for the R^2, we can see that $R^2 = 0.9979 \approx 1$. It is a great result since the value corresponds to the cosinus of the angle between the vector of the predicted value and the vector of the observed value, and the closer to 0 the angle gets, the better the model becomes.

```
Y <- tab$mau
Y_predict <- predict(modreg, tab)

plot(Y, Y_predict, main="Observed and predicted values", pch=19, col=rgb(0.27,0.4,0.68))
grid()
abline(a=0, b=1, col="red")
```

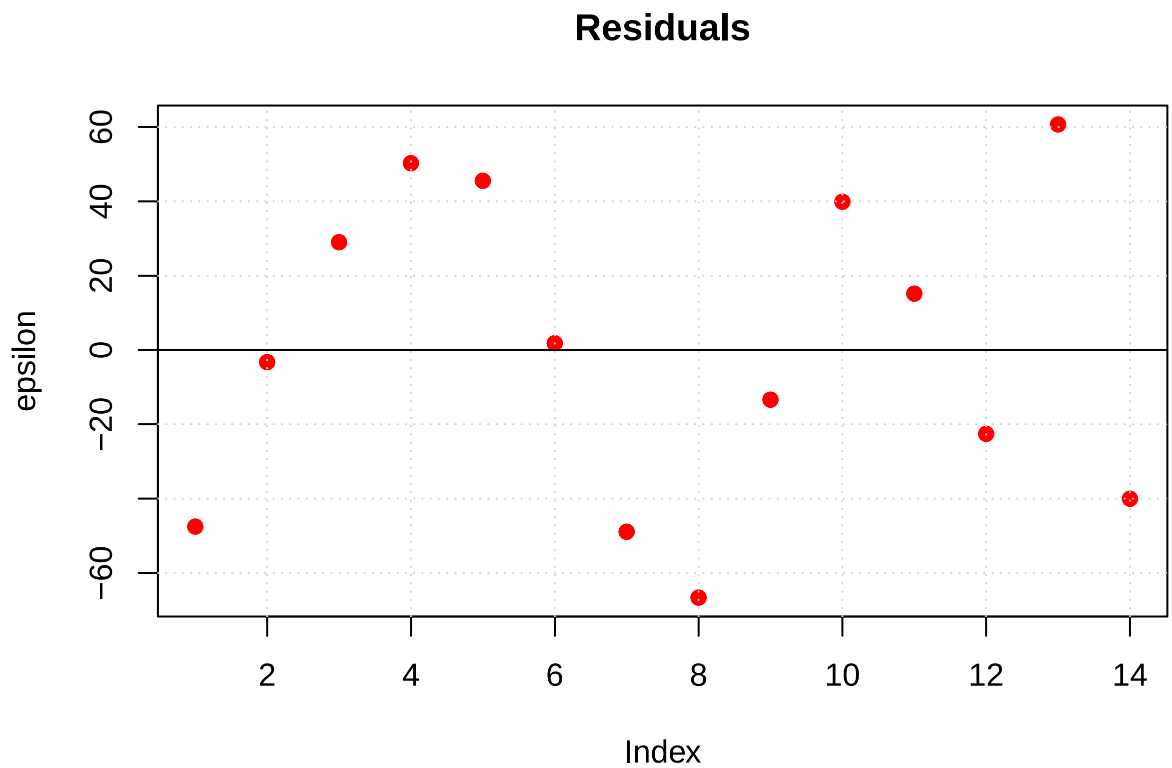


In the graph $(y, \hat{y})$ above, we can see that the plotted points are fairly close to the bisector, which indicates that the model is acceptable.

The graph below is a scatter plot of the residuals for every pair of $(y, \hat{y})$.

```
epsilon <- Y - Y_predict

plot(epsilon, main="Residuals", pch=19, col="red")
grid()
abline(a=0, b=0, col="black")
```



V. Medical data

```
tab <- read.table("diabetes.txt", header=TRUE, sep="\t")
```

The dimension of the data set:

```
dim(tab)
```

```
## [1] 442  11
```

The names of the variables:

```
names(tab)
```

```
##  [1] "AGE" "SEX" "BMI" "BP"  "S1"  "S2"  "S3"  "S4"  "S5"  "S6"  "Y"
```

The data set consists of $p = 10$ co-variables and one target variable ("Y"), and $n = 442$ observations.

We now try creating a linear model using the data set.

```r
modreg = lm(Y~., data=tab)
summary(modreg)
```

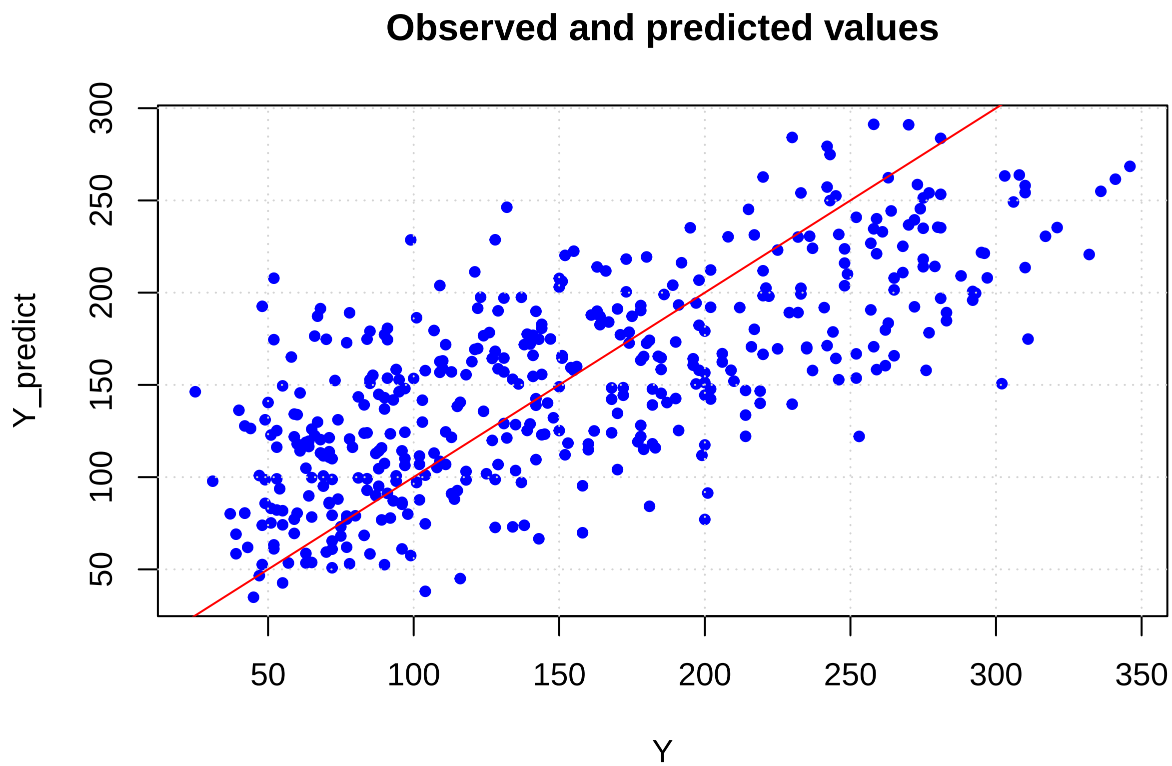
```
##
## Call:
## lm(formula = Y ~ ., data = tab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -155.827  -38.536   -0.228   37.806  151.353
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -334.56714   67.45462  -4.960 1.02e-06 ***
## AGE           -0.03636    0.21704  -0.168 0.867031
## SEX          -22.85965    5.83582  -3.917 0.000104 ***
## BMI            5.60296    0.71711   7.813 4.30e-14 ***
## BP             1.11681    0.22524   4.958 1.02e-06 ***
## S1            -1.09000    0.57333  -1.901 0.057948 .
## S2             0.74645    0.53083   1.406 0.160390
## S3             0.37200    0.78246   0.475 0.634723
## S4             6.53383    5.95864   1.097 0.273459
## S5            68.48312   15.66972   4.370 1.56e-05 ***
## S6             0.28012    0.27331   1.025 0.305990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 54.15 on 431 degrees of freedom
## Multiple R-squared:  0.5177, Adjusted R-squared:  0.5066
## F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16
```

With the `summary()` function, we can see that the variable "SEX", "BMI", "BP" and "S5" are the most significant variables of the model. We can also see that $R^2 = 0.5177$ which is quite far from 1, indicating that the current linear regression is not doing well.

Using the obtained linear model, we further predict the values of y and compare them with the observed values in the data set.

```r
Y <- tab$Y
Y_predict <- predict(modreg, tab)

plot(Y, Y_predict, col="blue", pch=20, main="Observed and predicted values")
grid()
abline(a=0, b=1, col="red")
```

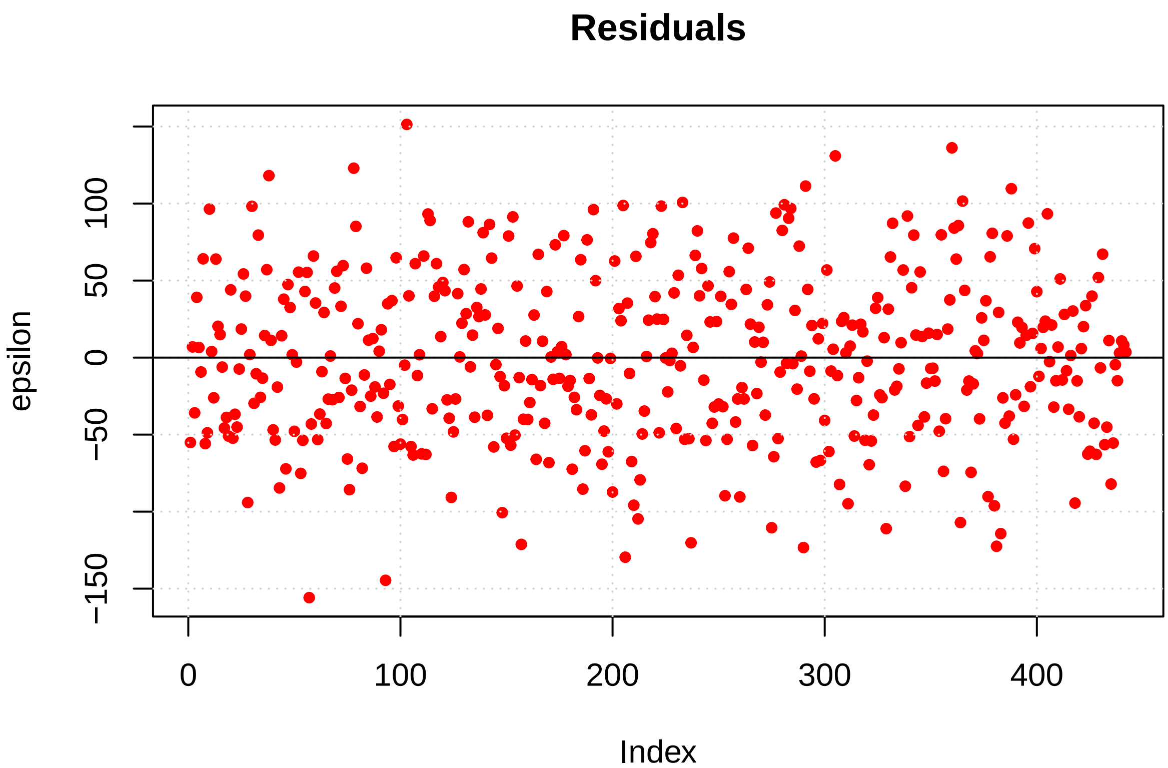


Based on the $(y, \hat{y})$ graph above, we can see that the plotted points are scattered around the bisector, with some points pretty far away from the bisector, rather that appearing on the line. This shows that the differences between the predicted values and the observed values are quite significant and thus the linear model is not good enough.

The scatter plot below shows the residual for every pair of $(y, \hat{y})$.

```
epsilon <- Y - Y_predict

plot(epsilon, col="red", pch=20, main="Residuals")
grid()
abline(a=0, b=0, col="black")
```


Residuals
epsilon
Index
−150
−50
0
50
100
0
100
200
300
400