Course 2 Section - 3.7 GOODNESS OF FIT

Jiaying Wu

16/10/2020

```r
#load library
library(tidyverse)
library(broom)
library(lubridate)
```

```r
# Read CO2 data and apply some pre-processing
CO2.spo <- read_csv(
  "https://raw.githubusercontent.com/datascienceprogram/ids_course_data/master/daily_merge_co2_spo.csv"
  col_names = c("date", "time", "day", "decdate", "n", "flg", "co2"), skip = 69) %>%
  filter(flg == 0) %>%
  mutate(date = ymd(date))

# Create variable day0 (a rescaling of day)
CO2.spo <- CO2.spo %>%
  mutate(day0 = day - min(day))
```

Give it a go

Explore the linear model of CO2 and with and without the quadratic term:

$co2 = \beta_0 + \beta_1 day_0 + \epsilon$

$co2 = \beta_0 + \beta_1 day_0 + \beta_2 day_0^2 + \epsilon$

- What is the $adjusted - R^2$ and BIC for both models?
- Which is the preferred model?

Model 1

```r
co2_mod1 <- lm(co2~day0, data=CO2.spo)
tidy(co2_mod1)
```

```
## # A tibble: 2 x 5
##   term         estimate std.error statistic p.value
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept) 302.      0.221         1367.       0
## 2 day0          0.00428 0.0000159      269.       0
```

```
glance(co2_mod1) %>%
  select(adj.r.squared, AIC)
```

```
## # A tibble: 1 x 2
##   adj.r.squared   AIC
##           <dbl> <dbl>
## 1         0.984 6193.
```

Model 2

```
co2_mod2 <- lm(co2~day0+I(day0^2), data=CO2.spo)
tidy(co2_mod2)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic p.value
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)  3.13e+2  1.01e- 1     3096.       0
## 2 day0         1.97e-3  1.80e- 5      109.       0
## 3 I(day0^2)    9.42e-8  7.13e-10      132.       0
```

```
glance(co2_mod2) %>%
  select(adj.r.squared, AIC)
```

```
## # A tibble: 1 x 2
##   adj.r.squared   AIC
##           <dbl> <dbl>
## 1         0.999 2898.
```

Since model 2 have the higher $adjusted - R^2$ and lower BIC, the model 2 is preferred.