More Tips and Tricks

Abhinav Anand, IIMB

2019/06/17

Scripts in R

Linear estimation in R

Generally a linear model takes the following form:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + u$$

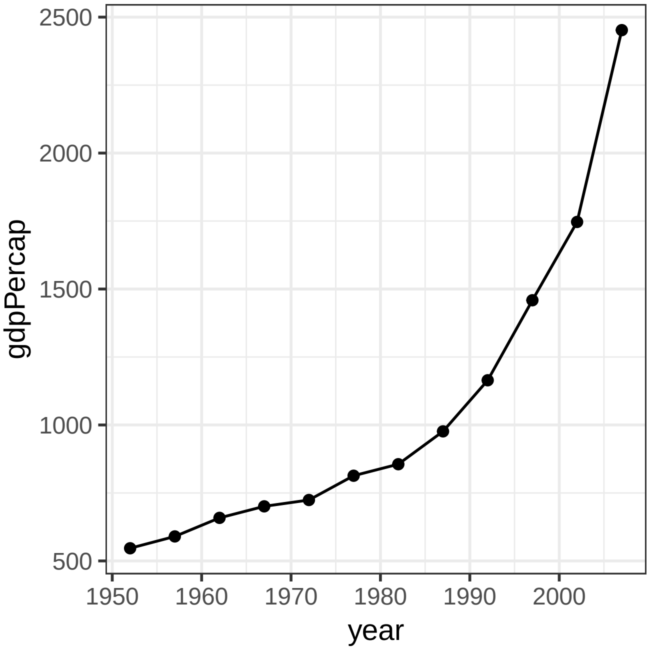
where $u_{n\times 1}$ is the error term. This setup corresponds to an overdetermined linear system of equations leading to a least squares solution:

$$\hat{\beta} = X^\top X^{-1} X^\top y$$

where the explanatory matrix $X_{m\times n}$ contains independent variables $x_1, \ldots, x_m$ as column vectors of size $n \times 1$.

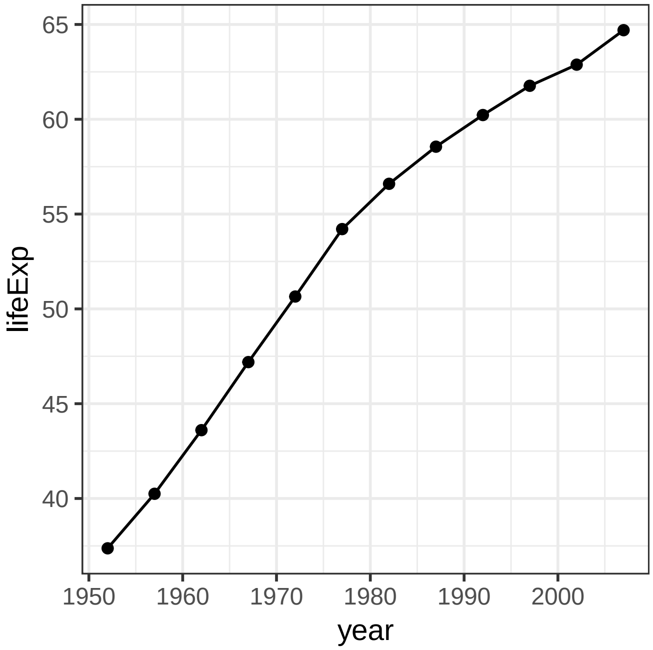
One of the strengths of R is the flexibility and support it offers for linear regression modeling. In order to illustrate it more fully, let us consider data for India in the `gapminder` dataset.

```
data_Ind <- gapminder::gapminder %>%
  dplyr::filter(country == "India")

ggplot(data_Ind, aes(year, gdpPercap)) +
  geom_point() +
  geom_line() +
  theme_bw()
```



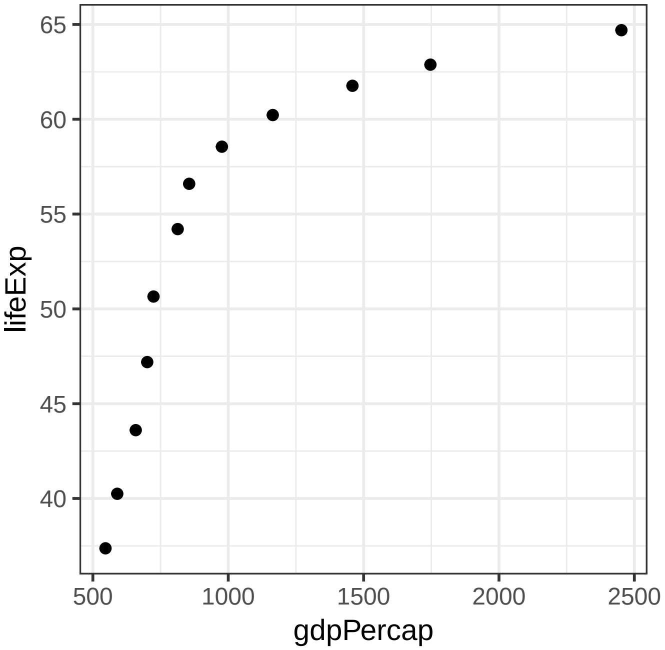
We see that there has been a large increase in GDP per capita in India. A similar trend is observed for life expectancy:

```
data_Ind <- gapminder::gapminder %>%
  dplyr::filter(country == "India")

ggplot(data_Ind, aes(year, lifeExp)) +
  geom_point() +
  geom_line() +
  theme_bw()
```



What about the relationship between the two? For example, (all else equal) does GDP per capita of India explain the life expectancy trends observed?

```
ggplot(data_Ind, aes(gdpPercap, lifeExp)) +
  geom_point() +
  theme_bw()
```



This suggests that the two variables share a positive relation. We can try to check this by means of a linear regression in the following way:

$$\text{life exp} = \beta_0 + \beta_1 \text{gdp percap} + u$$