Study on birth weight

Karim Saied (karim.saied@unil.ch)

November 12th, 2017

1. Question

The aim of the present study is to identify factors linked to a lower birth weight.

2. Dataset description

We have analysed a dataset from a study on birth weight and containing observations describing newborns and their mother (189 entries).

The newborns are described by the following variables:

- `bwt`: newborn weight (continous variable)
- `ptd`: preterm birth (two-levels factor)

The mother is described by the following variables:

- `lwt`: weight befor pregnancy (continuous variable)
- `age`: age (discret variable)
- `smoke`: smoking habits (two-levels factor)
- `ht`: known hypertension (two-levels factor)

First of all, no indication was provided regarding the unit of measurement of the weights. However, it is clear that this unit seems to differ between the newborns and the mothers. Indeed, the newborn's and mother's weight seem to be given in grams and pounds respectively. An unit conversion was performed in order to set both variables in kilograms.

Regarding the proportions within each factor (i.e. `smoke`, `ht`, `ptd`), it is clear that the dataset is biased:

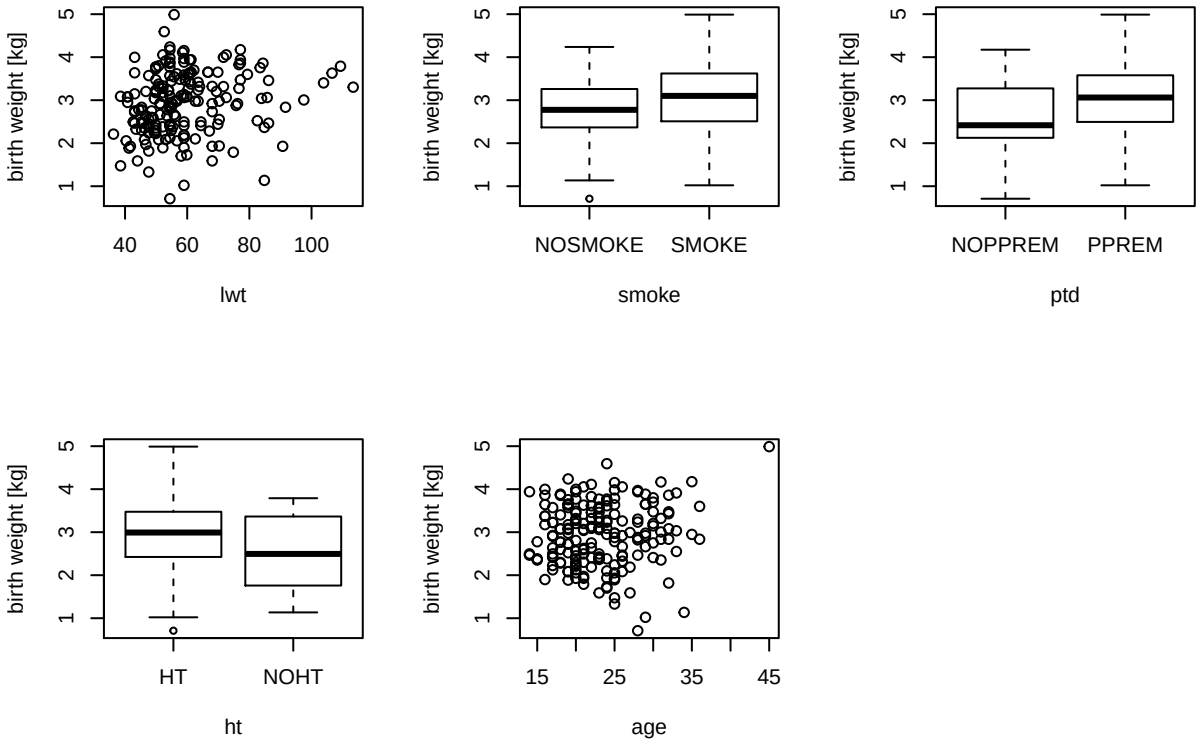
smoke	ht	ptd
NOSMOKE: 74	HT :177	NOPPREM: 30
SMOKE :115	NOHT: 12	PPREM :159

- 61% of the mothers smoke.
- 94% of the mothers have hypertension.
- 84% of the newborns are prematured.

These propotions are not representative of the population. According to the Centers for Disease Control and Prevention (CDC), the preterm birth rate in the US was 9.85% in 2016 and 15.1% of the adults were smokers in 2015. Regarding hypertension, one third of the US population is concerned (Merai et al., 2016).

3. Pre-analysis

Knowing the aim of the present analysis the newborn's weight (`bwt`) is considered as the response variable. Before any analysis, a data visualisation is performed.

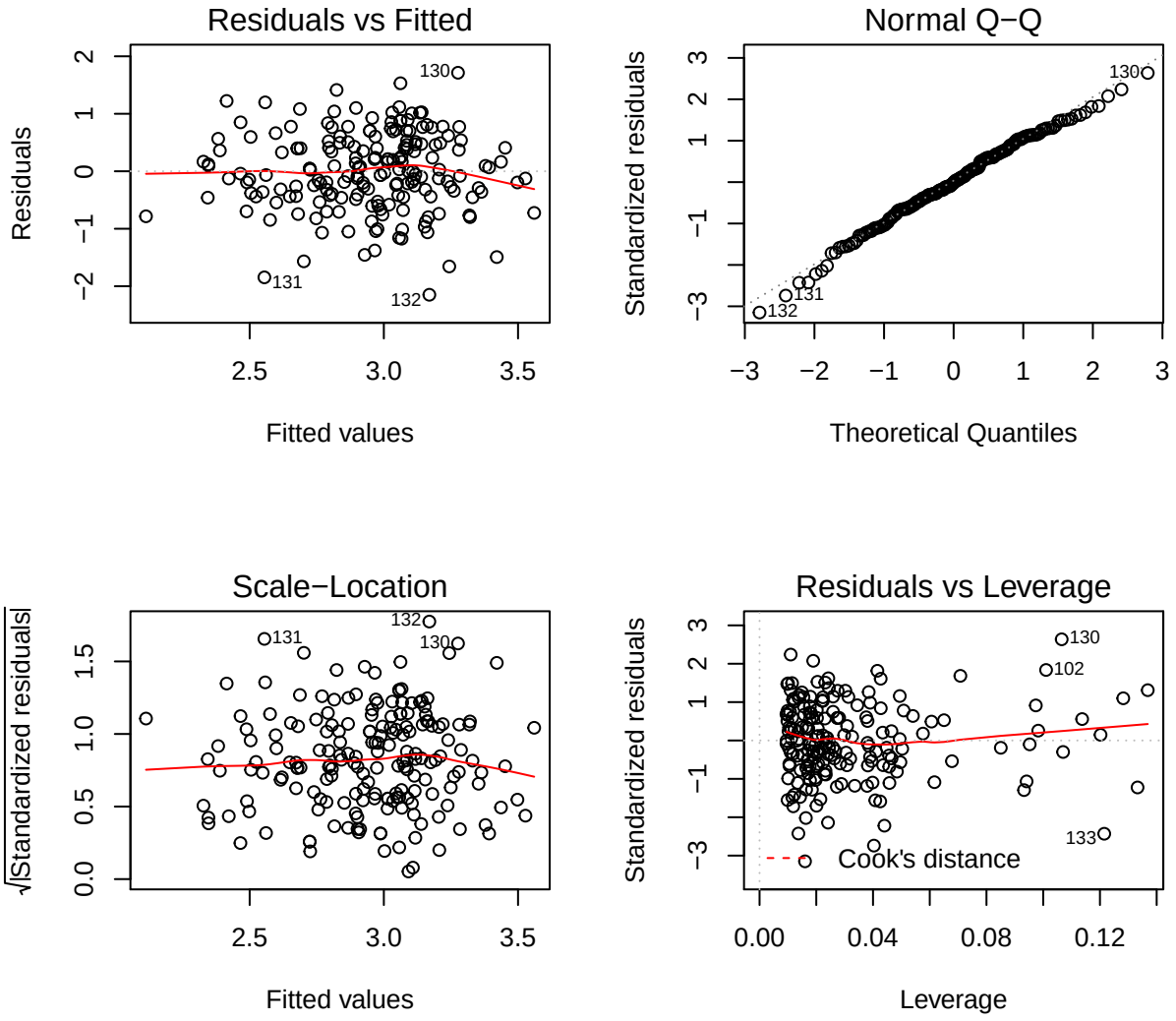


As shown in the boxplots, it seems to be a trend between the newborn's weight and the different factors. Indeed, mothers with hypertension as well as the ones that smoke tend to have heavier babies. Surprisingly, prematured newborns seem to be heavier than the others. This unexpected observation should lead us to consider the data with caution.

4. Statistical analysis

In order to assess the factors linked to a lower birth weight, a linear model is performed.

4.1. Final model



The graph in the top left corner (Residuals vs Fitted) shows no structure in the data as well as the graph in the bottom left corner. The top right graph shows that the residus are normally distributed. Taken together, these graphs show that a linear model is a suitable approach for the analysis.

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ptd + ht + age)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.14852 -0.43299 -0.00427  0.48872  1.71366
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.761936   0.310053   5.683 5.15e-08 ***
## lwt          0.010102   0.003829   2.638  0.00906 **
## smokeSMOKE   0.212918   0.104524   2.037  0.04309 *
## ptdPPREM     0.347253   0.141473   2.455  0.01504 *
## htNOHT      -0.559382   0.211720  -2.642  0.00895 **
## age          0.008681   0.009733   0.892  0.37364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6877 on 183 degrees of freedom
## Multiple R-squared:  0.1339, Adjusted R-squared:  0.1102
## F-statistic: 5.658 on 5 and 183 DF,  p-value: 7.078e-05
```

Note: several models have been tested in order to assess a possible interaction between the factors. The factors have been tested in pairs for an interaction because making a biological interpretation with several interactions between multiple factors would be complex and risky. The following model was one of the models tested for an interaction:

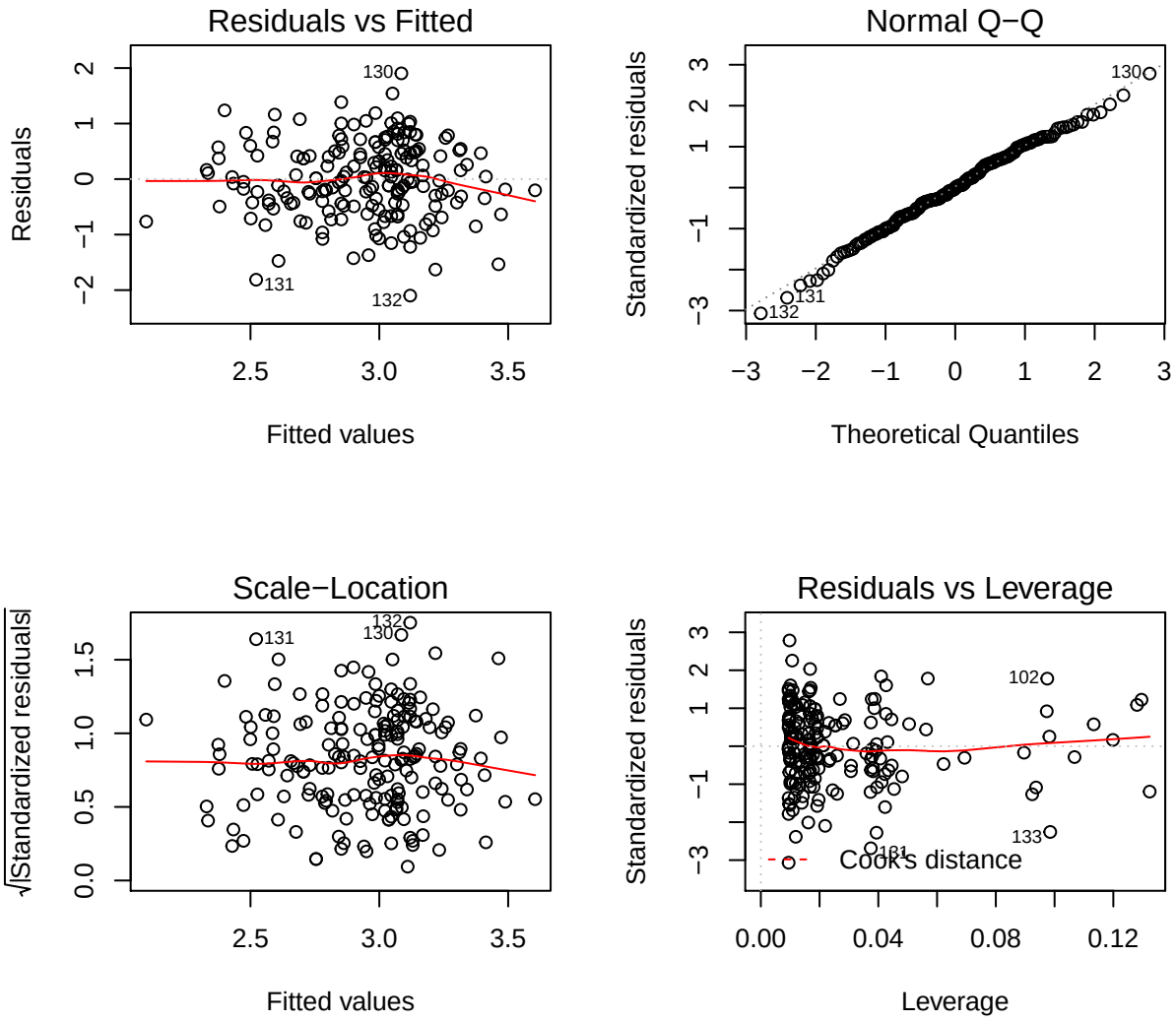
```
model_int <- lm(bwt~ht*smoke+ptd+lwt+age)
```

None of them highlighted a significant interaction between the factors. As a consequence the factors have been considered as independent in the final model.

The final model shows that the age do not influence significantly the newborn's weight (p-value = 0.374). Moreover, none of the models tested before (i.e. models tested for an interaction) shown significant influence of the age on newborn's weight. As a consequence, the age was excluded from the final model and a final model adjusted was performed.

4.2. Final model adjusted

The model was reshaped without the age factor.



The residus are still normally distributed and no structure is present in the data. The linear model without the age factor is analysed.

```
##
## Call:
## lm(formula = bwt ~ lwt + smoke + ptd + ht)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09885 -0.43842 -0.00606  0.47946  1.90442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.934496   0.242138   7.989 1.44e-13 ***
## lwt          0.010793   0.003748   2.880  0.00445 **
## smokeSMOKE   0.218471   0.104280   2.095  0.03754 *
## ptdPPREM     0.330463   0.140137   2.358  0.01942 *
## htNOHT      -0.571620   0.211157  -2.707  0.00743 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6873 on 184 degrees of freedom
## Multiple R-squared:  0.1301, Adjusted R-squared:  0.1112
## F-statistic: 6.881 on 4 and 184 DF,  p-value: 3.49e-05
```

5. Discussion and conclusion

The linear model shows that mother's weight befor pregnancy, smoking habit, hypertension and preterm delivery influence the newborn's weight significantly. On the contrary, the age had no significant influence. According to the statistical results and as observed previously with boxplots, prematured newborns are heavier than the others which is unexpected.

The factors are linked to a lower newborn's weight in the following way:

- A low mother's weight befor pregnancy (p-value = 0.004).
- A non-smoker habit (p-value = 0.038).
- A full term delivery (p-value = 0.019).
- No hypertension (p-value = 0.007).

Knowing unexpected results, the small sample size (189 entries) and the dataset bias, the results should be considered with caution. Although the current statistical analysis should be carried out on a bigger sample size, the experimental design of the original study should however be reshaped.