Week3-Exercises-Solutions

Exercise solutions

Week 3

The dataset lowbwt.csv was part of a study aiming to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams).

1 - Fit a linear model for the variable *bwt* (birth weight) using the covariate *age* (mother's age), evaluate the assumptions and interpret the results.

R code

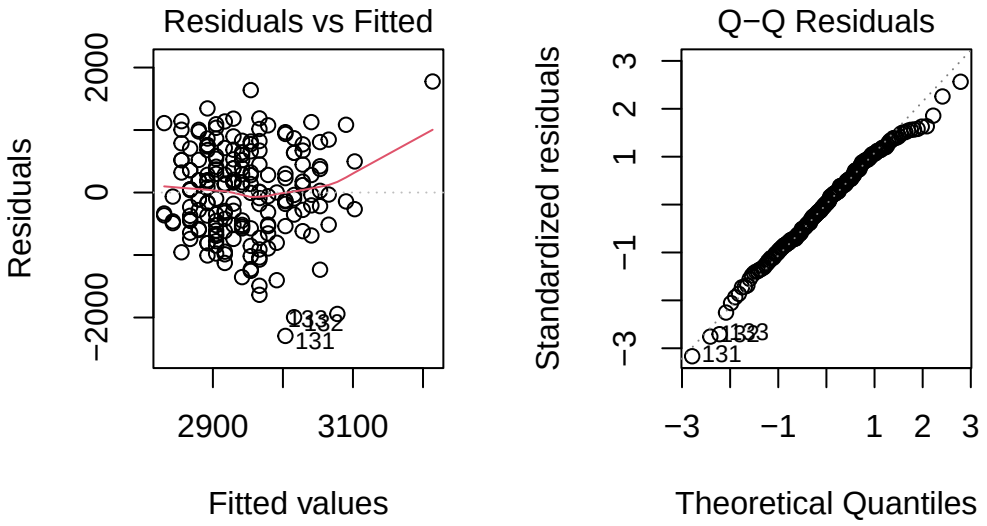
```
lowbwt <- read.csv("https://www.dropbox.com/scl/fi/ljqh7xojwidza0h1e7lg4/lowbwt.csv?rlkey=
lm1 <- lm(bwt~age, data=lowbwt)  #fit  the model
summary(lm1)
```

```
Call:
lm(formula = bwt ~ age, data = lowbwt)

Residuals:
     Min       1Q   Median       3Q      Max
-2294.53  -517.71    10.56   530.65  1776.27

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2657.33     238.80  11.128   <2e-16 ***
age            12.36      10.02   1.234    0.219
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 728 on 187 degrees of freedom
Multiple R-squared:  0.008076,  Adjusted R-squared:  0.002772
F-statistic: 1.523 on 1 and 187 DF,  p-value: 0.2188
```

```r
confint(lm1)
```

```
                 2.5 %     97.5 %
(Intercept) 2186.236438 3128.42869
age           -7.403527   32.13219
```

```r
par(mfrow=c(1,2))
plot(lm1, c(1,2)) #residuals vs fitted and distrib of the residuals
```



Stata code

```
clear
import delimited "https://www.dropbox.com/scl/fi/ljqh7xojwidza0h1e7lg4/lowbwt.csv?rlkey=st

reg bwt age
rvfplot
predict res_std, residuals /* calculate the residuals*/
qnorm res_std /* The normal quantile plot of the residuals*/
## (encoding automatically selected: UTF-8)
## (11 vars, 189 obs)
```

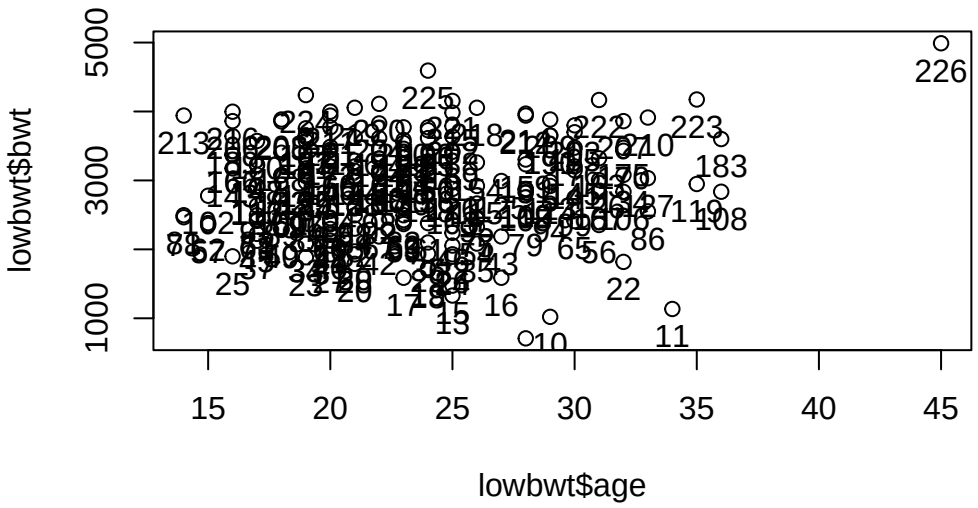
```
##
##
##       Source |       SS           df       MS      Number of obs   =       189
## -------------+----------------------------------   F(1, 187)       =      1.52
##        Model |  806926.913         1  806926.913   Prob > F        =    0.2188
##     Residual |  99110125.7       187  530000.672   R-squared       =    0.0081
## -------------+----------------------------------   Adj R-squared   =    0.0028
##        Total |  99917052.6       188  531473.684   Root MSE        =    728.01
##
## ------------------------------------------------------------------------------
##          bwt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
## -------------+----------------------------------------------------------------
##          age |   12.36433   10.02055     1.23   0.219    -7.403527    32.13219
##        _cons |   2657.333    238.804    11.13   0.000     2186.236    3128.429
## ------------------------------------------------------------------------------
```

We have fitted a linear regression for birth weight using mother's age. The first residual plot (residuals vs fitted) does not show evidence of departure from the linearity assumption or violation of the homoscedasticity assumption as the dispersion seems constant along the x-axis. The q-q plot indicates that the residuals appear to follow a normal distribution (a bit of departure at one of the tails).

From the regression output, the intercept 2657, which is the expected birth weight for a mother with an age of zero. The coefficient for age is 12, which means that for every one year increase in the mother's age, the birth weight is expected to increase 12 grams in average (95%CI: [-7, 32]). The R-squared value for this model is 0.008. This means that 0.8% of the variability in birth weight can be explained by the mother's age. This result does not provide evidence of a linear relation between birth weight and mother's age.

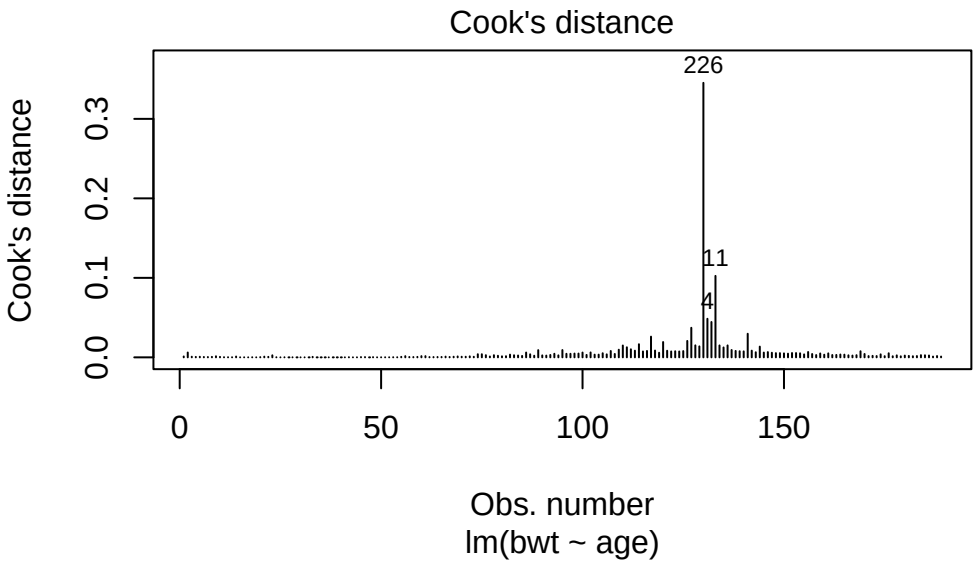
2 - Evaluate potential outliers and influential observations. How would the results change if you excluded this/these observation(s)?

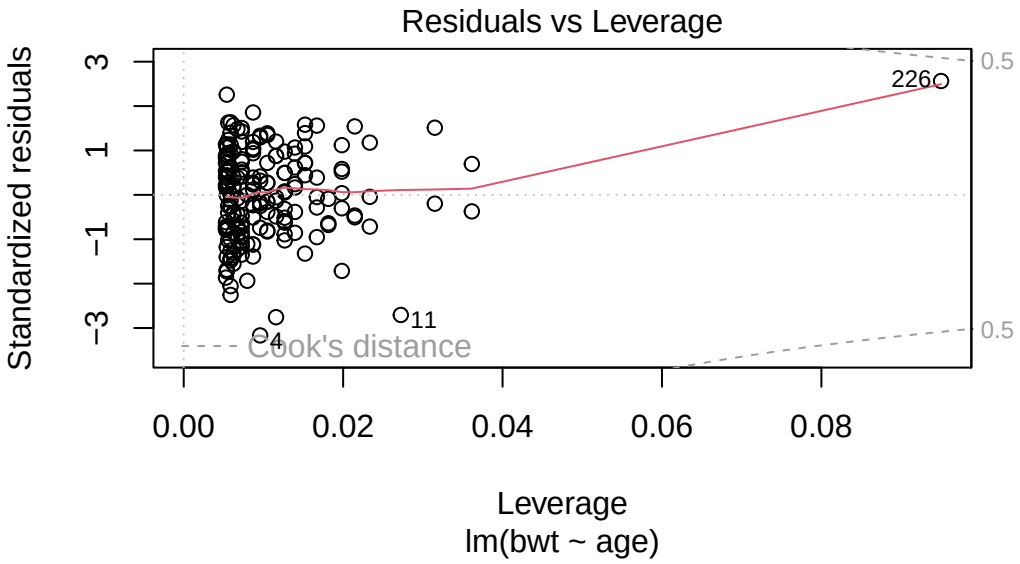
```
plot(lowbwt$age, lowbwt$bwt)
text(lowbwt$age, lowbwt$bwt, labels = lowbwt$id, pos = 1) #adds ids
```



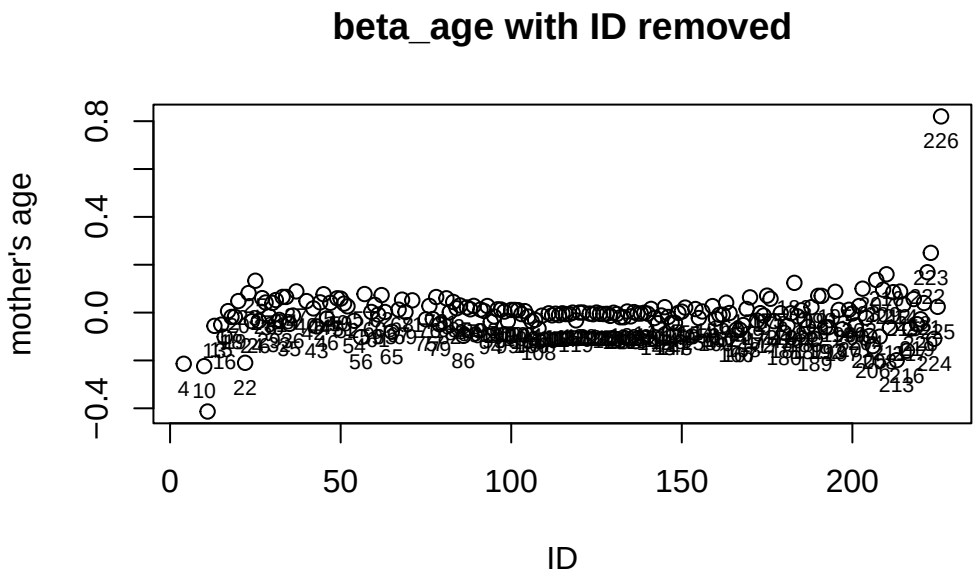
From the scatter plot, we can recognise the observation with id=226 has a value for age far from the other observations. Let's look at some plots with influence measures:

```
#Cook's distance
#the option labels.id shows the label from
#the variable id rather than the row number
plot(lm1, c(4,5), labels.id=lowbwt$id)
```


Cook's distance
226
11
4
Cook's distance
0.0
0.1
0.2
0.3
0
50
100
150
Obs. number
lm(bwt ~ age)


Residuals vs Leverage
226
11
4
Cook's distance
0.5
0.5
Standardized residuals
−3
−1
1
3
0.00
0.02
0.04
0.06
0.08
Leverage
lm(bwt ~ age)

```
#Plotting the change in the beta for age
#when each observation is removed
inf <- influence.measures(lm1)
plot(lowbwt$id, inf$infmat[,2], xlab="ID", ylab="mother's age")
text(lowbwt$id, inf$infmat[,2], labels = lowbwt$id, pos = 1, cex=.7) #adds ids
title("beta_age with ID removed")
```




```
clear
import delimited "https://www.dropbox.com/s/r06u1l1cjrvcpl1/lowbwt.csv?dl=1"
reg bwt age
dfbeta
/*leverage plot /*
lvr2plot

/*plotting the change in the beta for age  (dfbeta)
when each observation is removed/*
twoway (scatter _dfbeta_1 id, sort mlabel(id))
## server refused to send file
## could not open url
## r(603);
##
```

```
## r(603);
```

After reviewing the plots above, it appears that both observation 11 and 226 have a significant impact on the estimation. As a general rule, it is not advisable to remove observations from the analysis without a strong justification, and any such decision should be well-justified. However, in this instance, observation 226 was made on a mother who is nearly 10 years older than the older mothers in the remaining cohort. This factor may suggest that this mother belongs to a different cohort, and we could argue for her exclusion from the analysis. Regardless, it would have been preferable to anticipate this issue during the study's design phase, for instance, by setting age$<$40 years old as an inclusion criterion.