Introduction to Medical Statistics 2026 Exercise 6 – Simple Linear Regression Data Analysis and Model Diagnostics

Nguyen Lam Vuong and Biostatistics team

2026-03-23

Exercise i) (Simple Linear Regression – HIV-negative CM patients)

As in the exercises of day 1, we use the dataset cmTbmData.csv containing information on 201 patients with meningitis from 4 different patient groups. However, for this session, we will restrict attention to the 49 HIV-negative patients with cryptococcal meningitis. We will examine how well blood white cell count can predict CSF white cell count in this group.

a) Import the dataset cmTbmData.csv and create a new data.frame cm.hivneg which contains HIV-negative patients with cryptococcal meningitis only.

```
cmTbm <- read.csv("https://raw.githubusercontent.com/oucru-biostats/IntroductionToBios
summary(cmTbm)
```

```
      code          hiv         diagnosis     group              groupLong
 BK01   :  1   Min.   :0.0000   CM :100   Min.   :1.000   HIV neg - CM :49
 BK02   :  1   1st Qu.:0.0000   TBM:101   1st Qu.:1.000   HIV neg - TBM:43
 BK03   :  1   Median :1.0000             Median :2.000   HIV pos - CM :51
 BK04   :  1   Mean   :0.5423             Mean   :2.413   HIV pos - TBM:58
 BK05   :  1   3rd Qu.:1.0000             3rd Qu.:3.000
 BK06   :  1   Max.   :1.0000             Max.   :4.000
 (Other):195
      age             sex            bldwcc          bldneut
```

```
 Min.   :15.00   Min.   :1.000   Min.   : 0.540   Min.   : 0.310
 1st Qu.:23.00   1st Qu.:1.000   1st Qu.: 5.985   1st Qu.: 4.445
 Median :28.00   Median :1.000   Median : 8.335   Median : 6.665
 Mean   :32.08   Mean   :1.259   Mean   : 9.484   Mean   : 7.710
 3rd Qu.:37.00   3rd Qu.:2.000   3rd Qu.:11.700   3rd Qu.: 9.620
 Max.   :78.00   Max.   :2.000   Max.   :26.700   Max.   :24.240
 NA's   :1                       NA's   :3        NA's   :9
     bldlym           csfwcc          csfneut           csflym
 Min.   :0.0500   Min.   :   0.0   Min.   :   0.00   Min.   :   0.00
 1st Qu.:0.5600   1st Qu.:  46.5   1st Qu.:   8.54   1st Qu.:  29.50
 Median :0.8000   Median : 142.0   Median :  40.00   Median :  72.21
 Mean   :0.9326   Mean   : 367.1   Mean   : 235.85   Mean   : 135.65
 3rd Qu.:1.1950   3rd Qu.: 453.5   3rd Qu.: 186.67   3rd Qu.: 182.85
 Max.   :7.8850   Max.   :3960.0   Max.   :3647.20   Max.   :1097.60
 NA's   :10       NA's   :6        NA's   :11        NA's   :10
```

```r
cm.hivneg  <- subset(cmTbm, groupLong=="HIV neg - CM")
```

b) Perform a linear regression with CSF white cell count as the outcome (response variable) and blood white cell count as a covariable (explanatory variable) and interpret the output. Calculate the 95% confidence intervals for the regression coefficients. Use the functions lm, summary, and confint. What do you conclude from the model results?

Add the fitted regression line to the scatterplot. (You can use the GUI in the ggplotgui package to obtain the scatterplot with the fitted regression line.) By looking at the plot, do you think that the model assumptions are fulfilled?

```r
fit <- lm(csfwcc ~ bldwcc, data = cm.hivneg)

# Summarize the regression result
summary(fit)
```

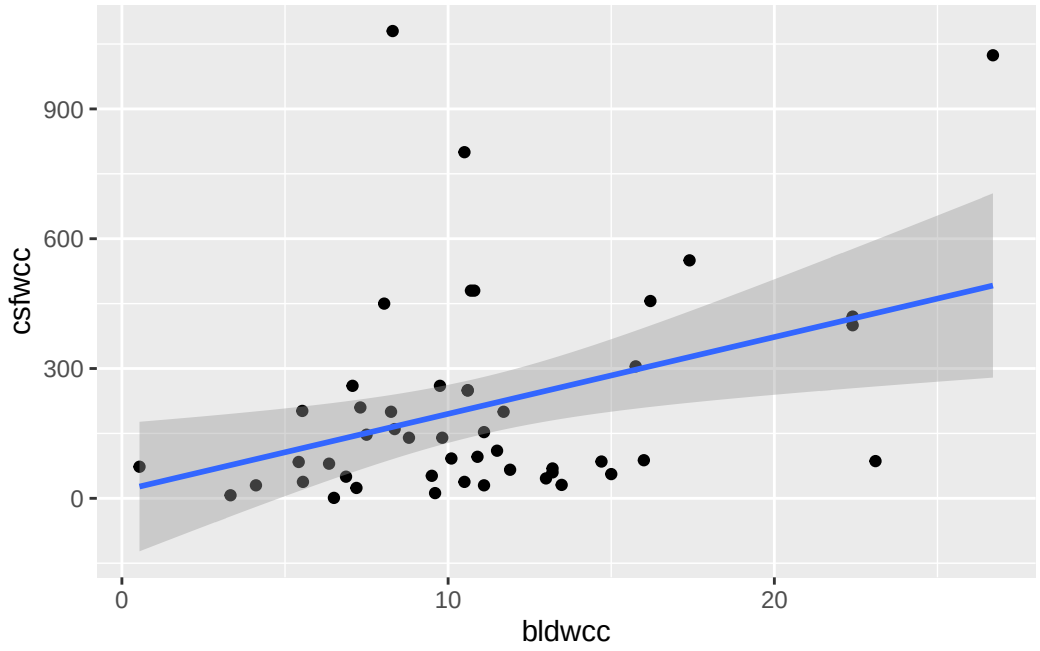
```
Call:
lm(formula = csfwcc ~ bldwcc, data = cm.hivneg)

Residuals:
    Min      1Q  Median      3Q     Max
-341.81 -134.30  -50.36   45.82  915.01
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.594     77.533   0.227  0.82147
bldwcc        17.758      6.398   2.776  0.00788 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 229.3 on 47 degrees of freedom
Multiple R-squared:  0.1408,    Adjusted R-squared:  0.1226
F-statistic: 7.705 on 1 and 47 DF,  p-value: 0.00788
```

```r
confint(fit) # Confidence intervals
```

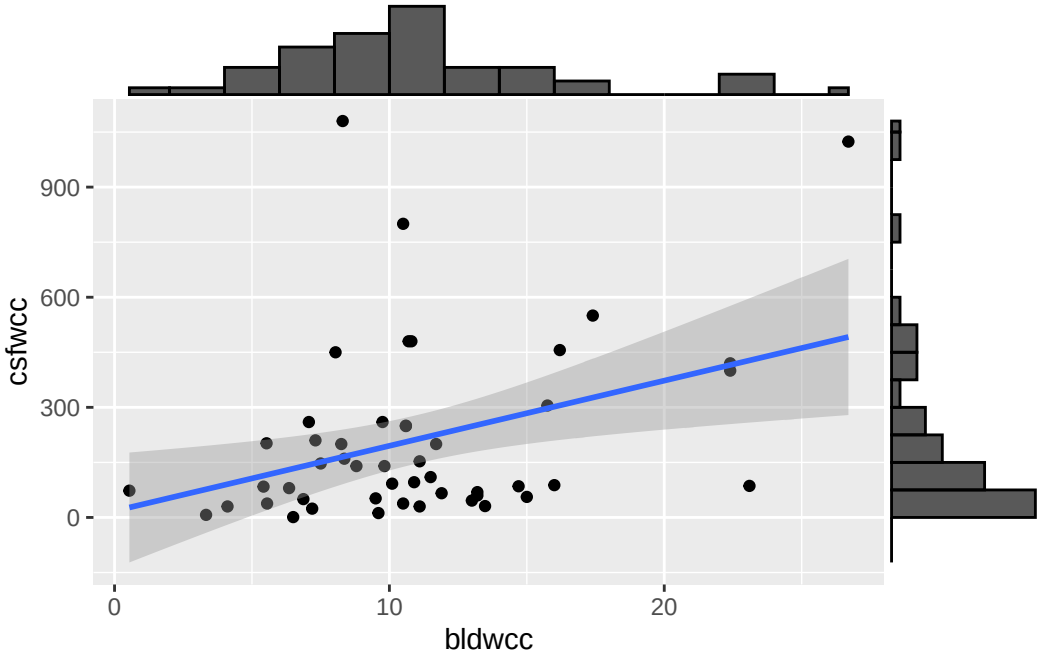
```
                 2.5 %    97.5 %
(Intercept) -138.383258 173.57109
bldwcc         4.888117  30.62858
```

```r
# Draw the scatterplot again (and add a regression line)
ggplot(cm.hivneg, aes(bldwcc, csfwcc)) +
  geom_point() +
  geom_smooth(method = "lm")
```



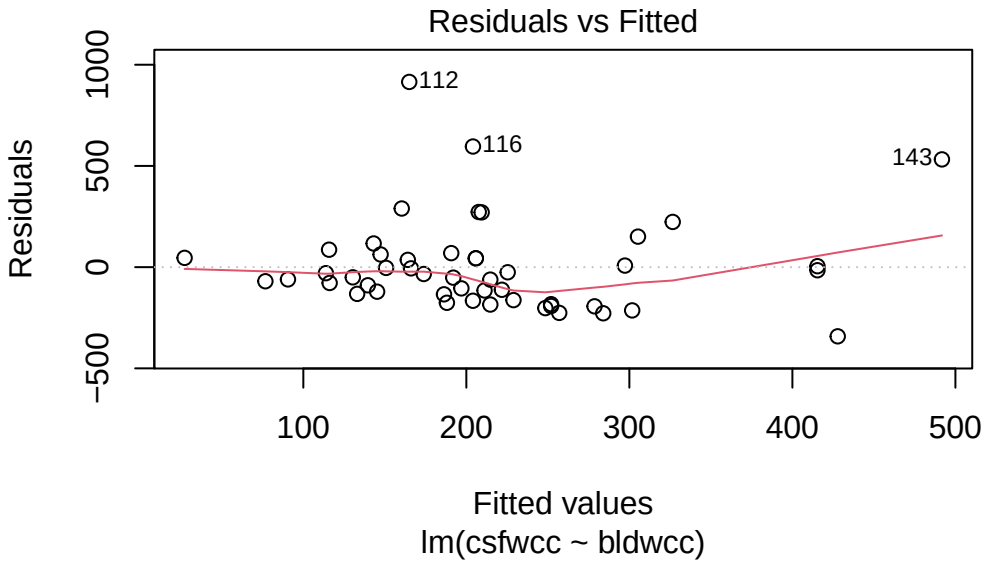
Answer: The test that the parameter for bldwcc is zero has a p-value of 0.008. Hence, assuming that the model assumptions are correct, there is a relation between blood and CSF white cell count. However, there are some points way above the fitted line, while this is not the case for points below the fitted line. Hence, the residuals do not follow a normal distribution.

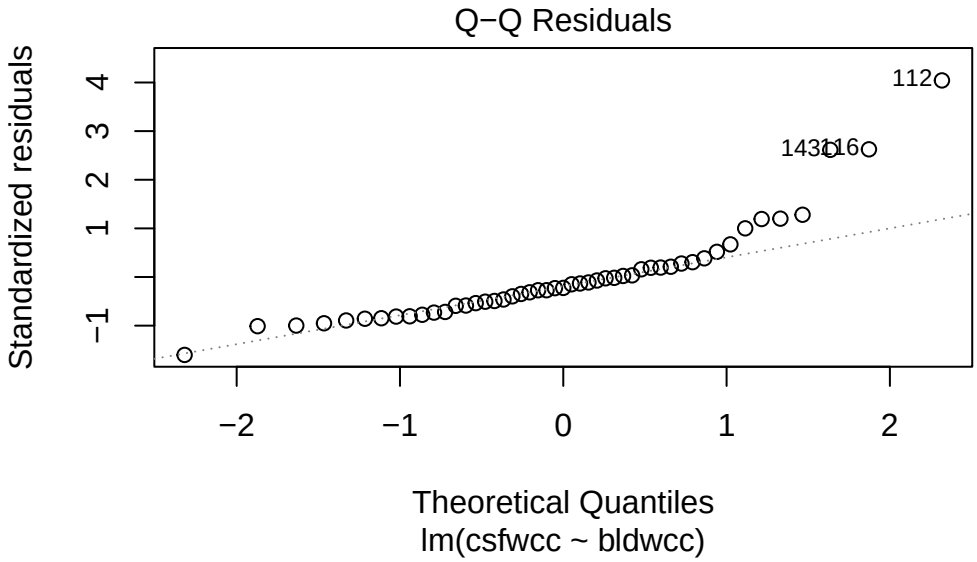
```
library(ggExtra)
# the ggExtra package allows to create the histograms (or boxplots) alongside the scat
# you may have to install it first
p <- ggplot(cm.hivneg, aes(bldwcc,csfwcc)) + geom_point() + geom_smooth(method="lm")
ggMarginal(p, type = "histogram", xparams=list(binwidth=2, boundary=0),
           yparams=list(binwidth=75, boundary=0))
```

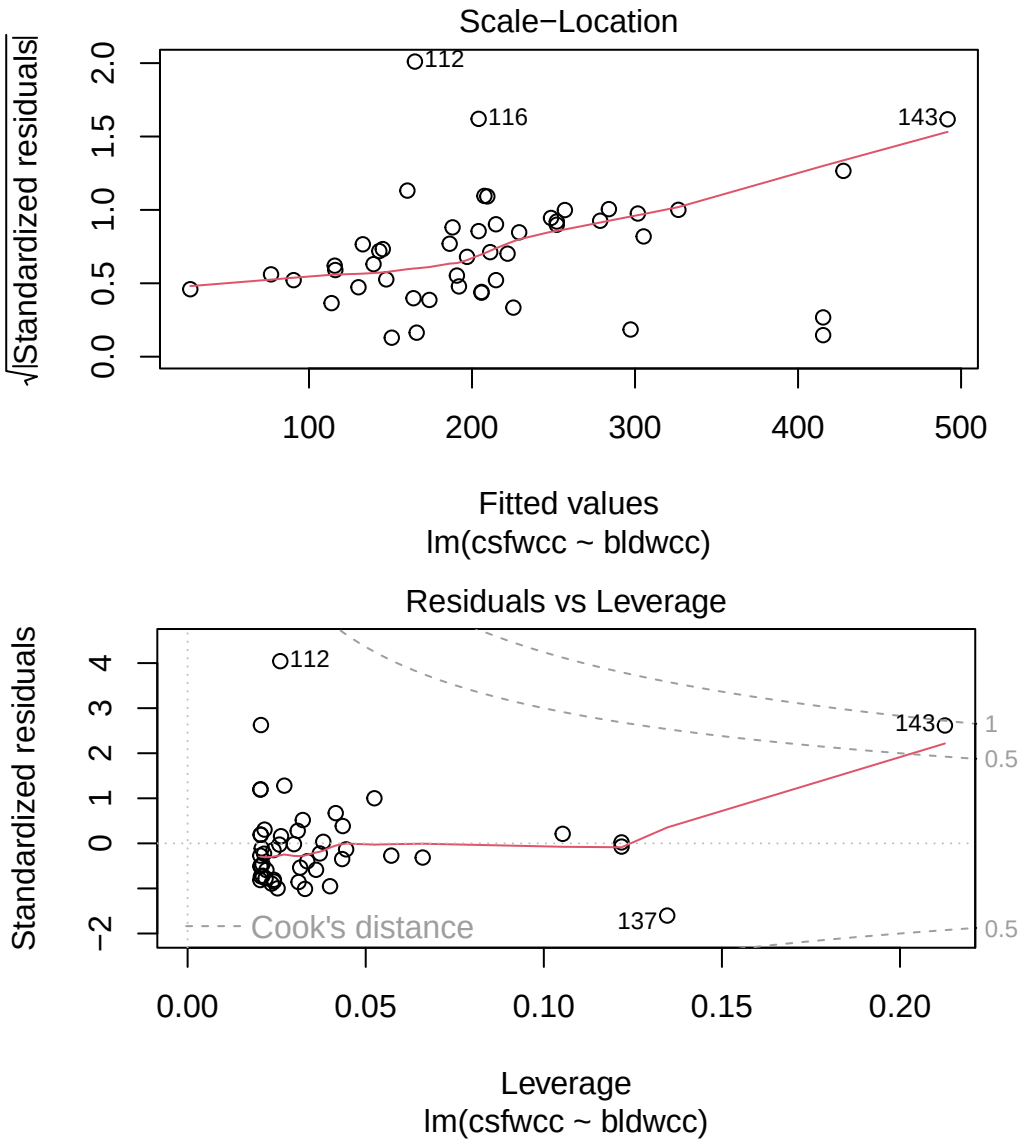


c) Perform diagnostic plots for the fitted model using plot(fit). Interpret the residuals. Do they indicate any problems regarding the assumptions of the linear regression model? (Some further explanation of diagnostic plots in R can be obtained at http://data.library.virginia.edu/diagnostic-plots/ (https://easystats.github.io/performance/))

```
plot(fit)
```


Residuals vs Fitted
Residuals
1000
500
0
−500
112
116
143
100
200
300
400
500
Fitted values
lm(csfwcc ~ bldwcc)

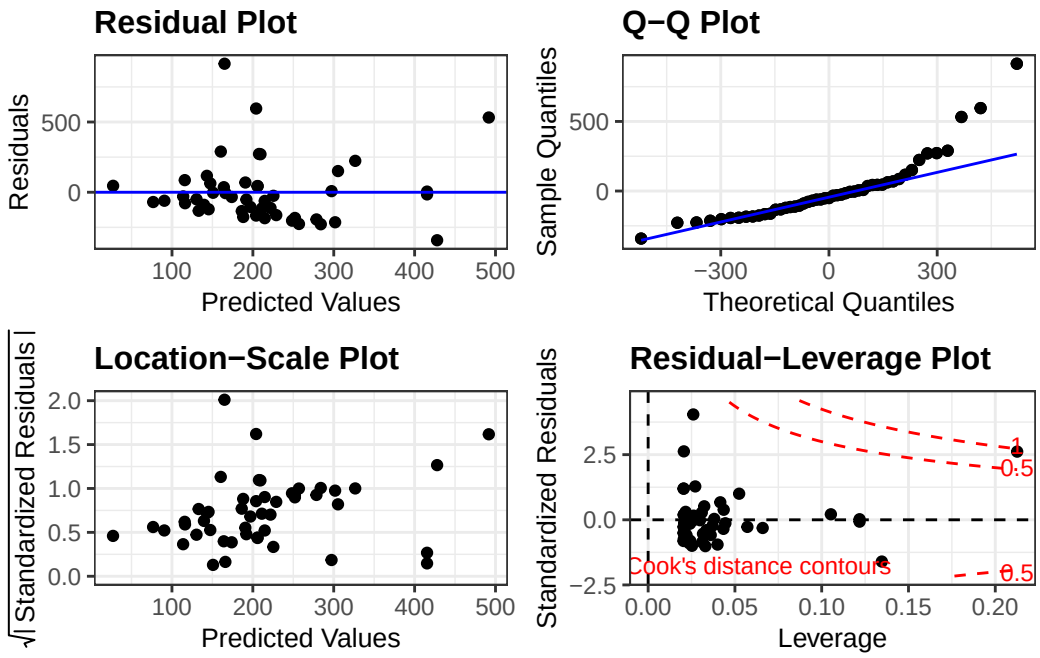

Q−Q Residuals
Standardized residuals
4
3
2
1
−1
112
143
116
−2
−1
0
1
2
Theoretical Quantiles
lm(csfwcc ~ bldwcc)



Answer: Also the residual plots don't look very good. There are some residuals that are much larger than 0. From the Q-Q plot we conclude that the residuals don't follow a normal distribution.

An alternative is to use the *resid_panel* function frim the **ggResidpanel** package. Using the argument *plot="R"* gives the same four plots.

```
library(ggResidpanel)
resid_panel(fit, plot="R")
```




```
library(performance)
```

```
Warning: package 'performance' was built under R version 4.4.3
```

```
library(see)
```

```
Warning: package 'see' was built under R version 4.4.3
```

```
library(patchwork)
```

```
Warning: package 'patchwork' was built under R version 4.4.3
```

```
# checking model assumptions
check_model(fit)
```

d) Create two new variables log10.bldwcc and log10.csfwcc containing log10-transformed values of the original data and then perform steps b) and c) again for the log-transformed variables. What do you conclude?

```r
# First check if there is any 0 in each wcc by seeing if the lowest value is 0
summary(cm.hivneg$bldwcc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.54    7.50   10.50   10.98   13.20   26.70
```

```r
summary(cm.hivneg$csfwcc)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0    56.0   110.0   212.7   260.0  1080.0
```

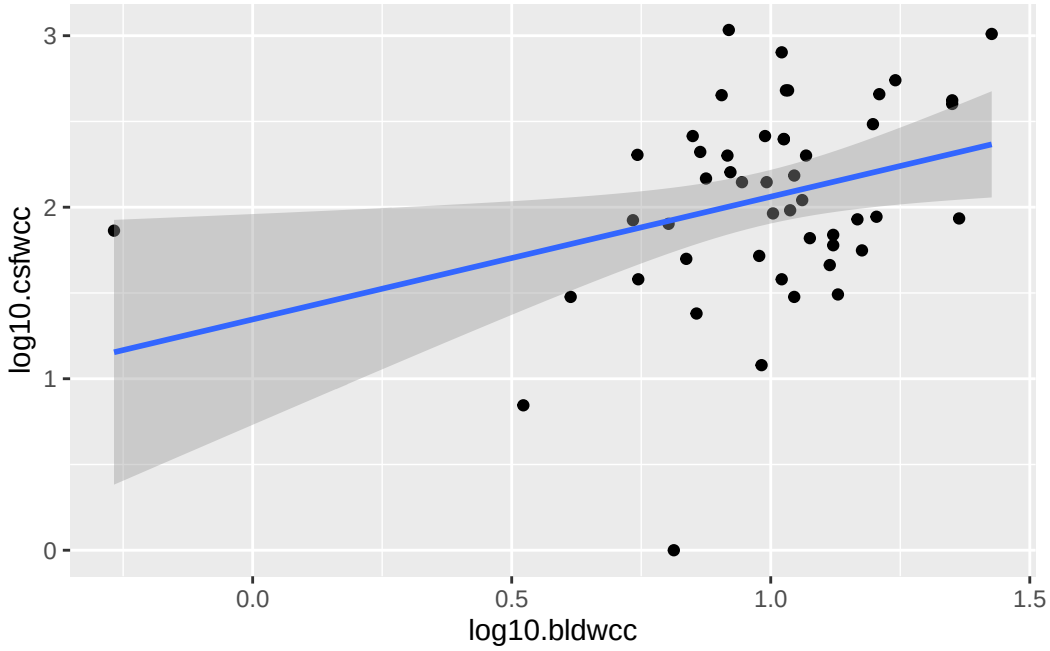
```r
# Create new variables with log10-transformed values
cm.hivneg$log10.bldwcc <- log10(cm.hivneg$bldwcc)
cm.hivneg$log10.csfwcc <- log10(cm.hivneg$csfwcc)

# Repeat steps b) - c)
fit.log <- lm(log10.csfwcc ~ log10.bldwcc, data = cm.hivneg)
summary(fit.log)
```

```
Call:
lm(formula = log10.csfwcc ~ log10.bldwcc, data = cm.hivneg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.92746 -0.30877  0.09044  0.35827  1.03000

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.3457     0.3054   4.406 6.05e-05 ***
log10.bldwcc   0.7156     0.3003   2.383   0.0213 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5428 on 47 degrees of freedom
Multiple R-squared:  0.1078,    Adjusted R-squared:  0.08881
F-statistic: 5.678 on 1 and 47 DF,  p-value: 0.02127
```

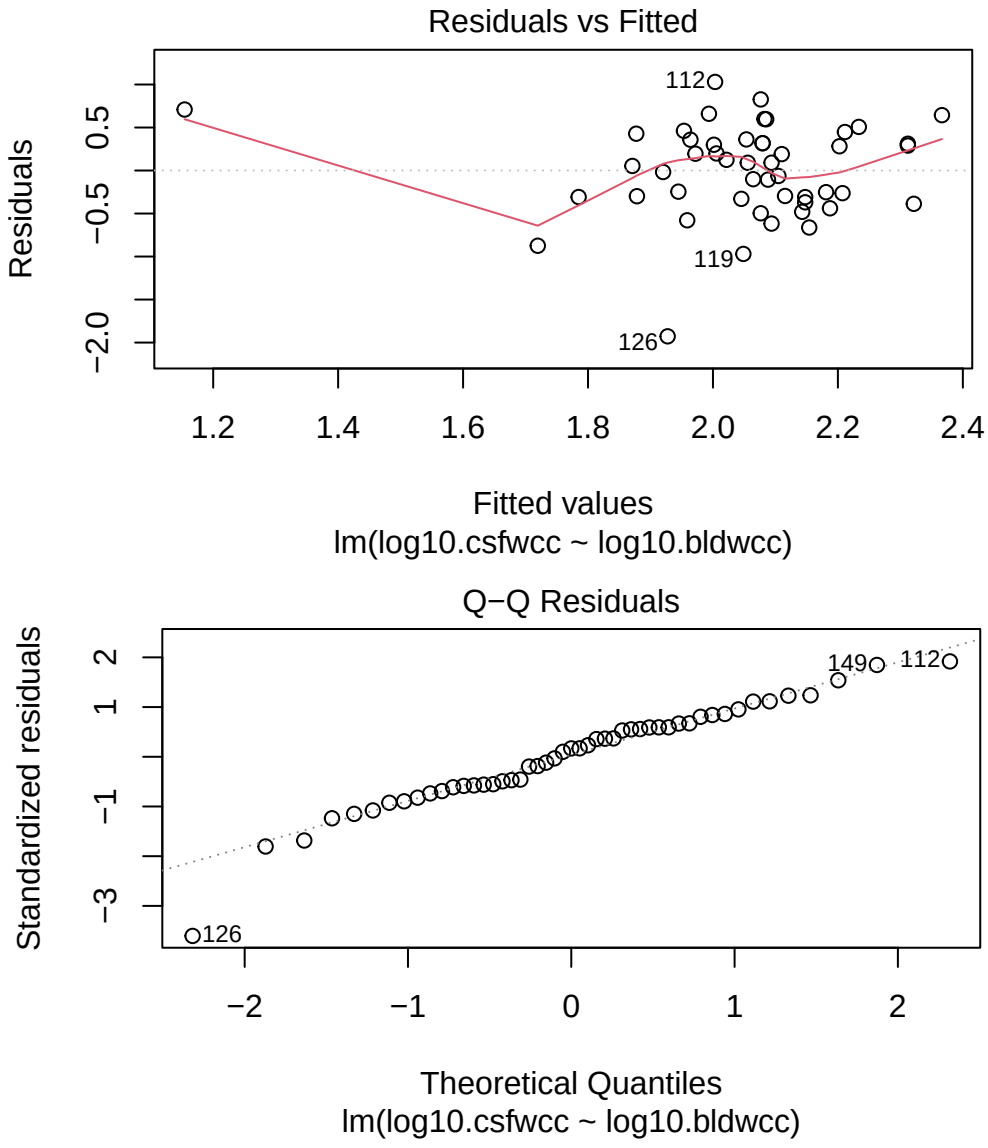
```r
confint(fit.log)
```

```
                2.5 %   97.5 %
(Intercept)  0.7313394 1.960120
log10.bldwcc 0.1114733 1.319739
```

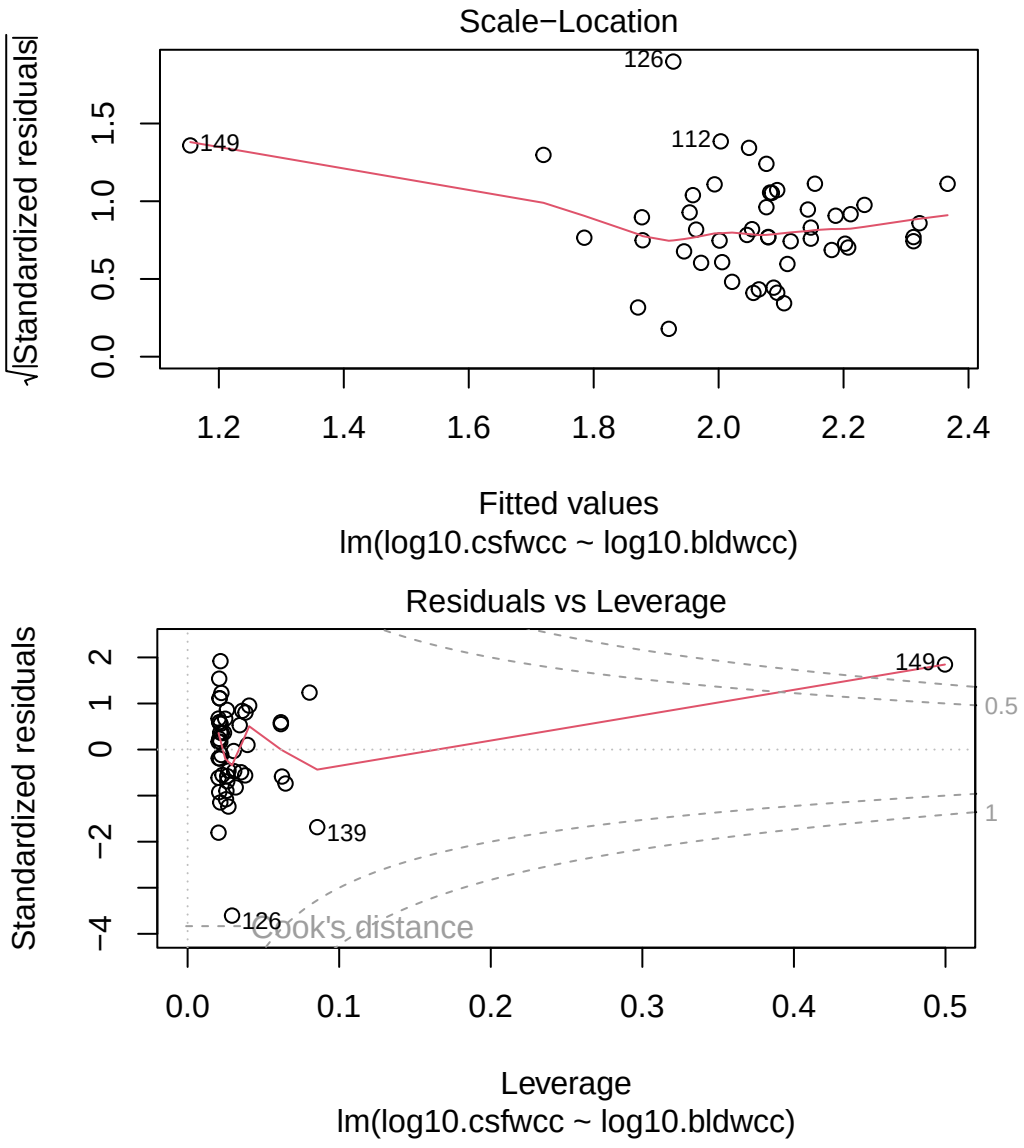
```r
ggplot(cm.hivneg, aes(log10.bldwcc, log10.csfwcc)) +
  geom_point() +
  geom_smooth(method = "lm")
```




```r
plot(fit.log)
```

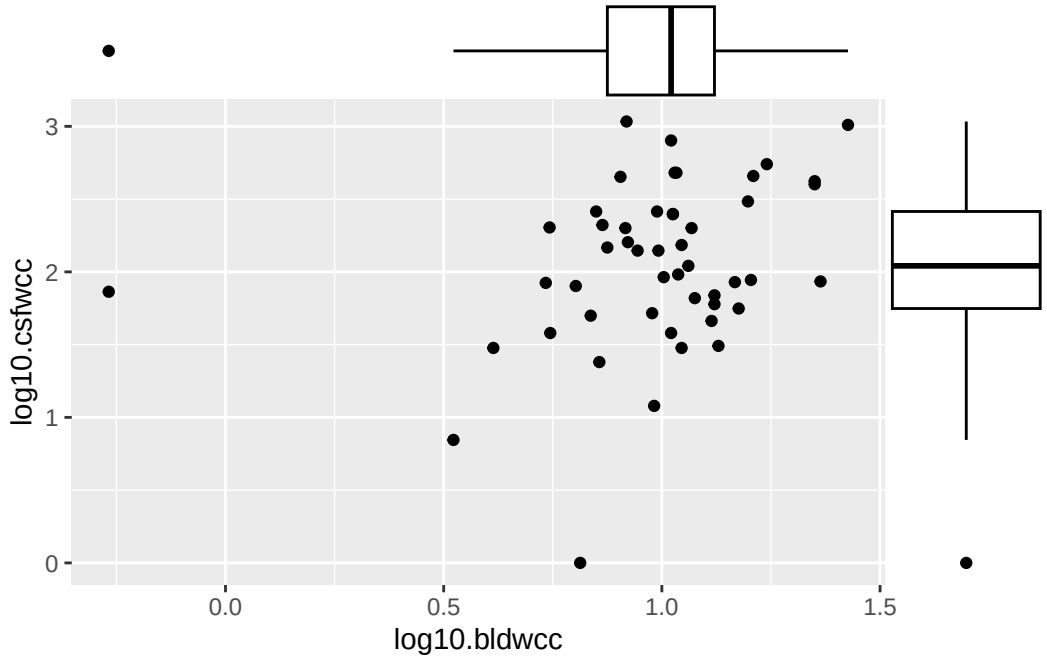

Residuals vs Fitted
Residuals
0.5
−0.5
−2.0
112
119
126
1.2
1.4
1.6
1.8
2.0
2.2
2.4
Fitted values
lm(log10.csfwcc ~ log10.bldwcc)

Q−Q Residuals
Standardized residuals
2
1
−1
−3
149
112
126
−2
−1
0
1
2
Theoretical Quantiles
lm(log10.csfwcc ~ log10.bldwcc)

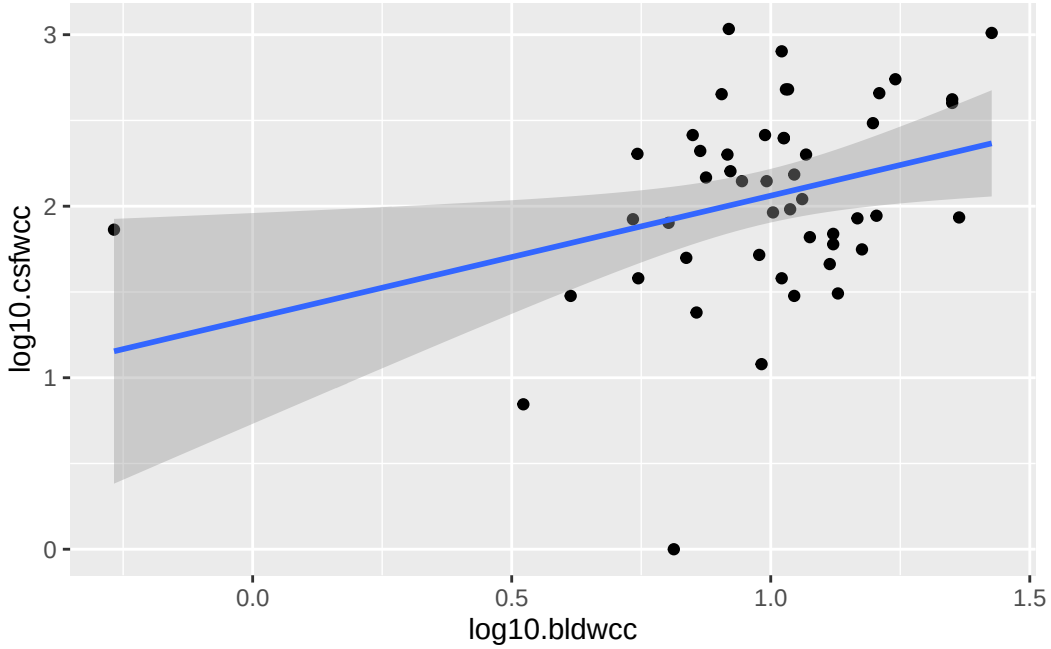


Answer d: 1. Diagnostic plots look better though there are some suspicious points. 2. Q-Q plot: The distribution looks more normal, except for one outlier. 3. The "Residuals vs Leverage" plot suggests that there's one individual that may have a large impact on the parameter estimates. 4. There is still a fairly strong suggestion that white blood cell count relates to CSF white cell count, although the p-value has gone up quite a bit.

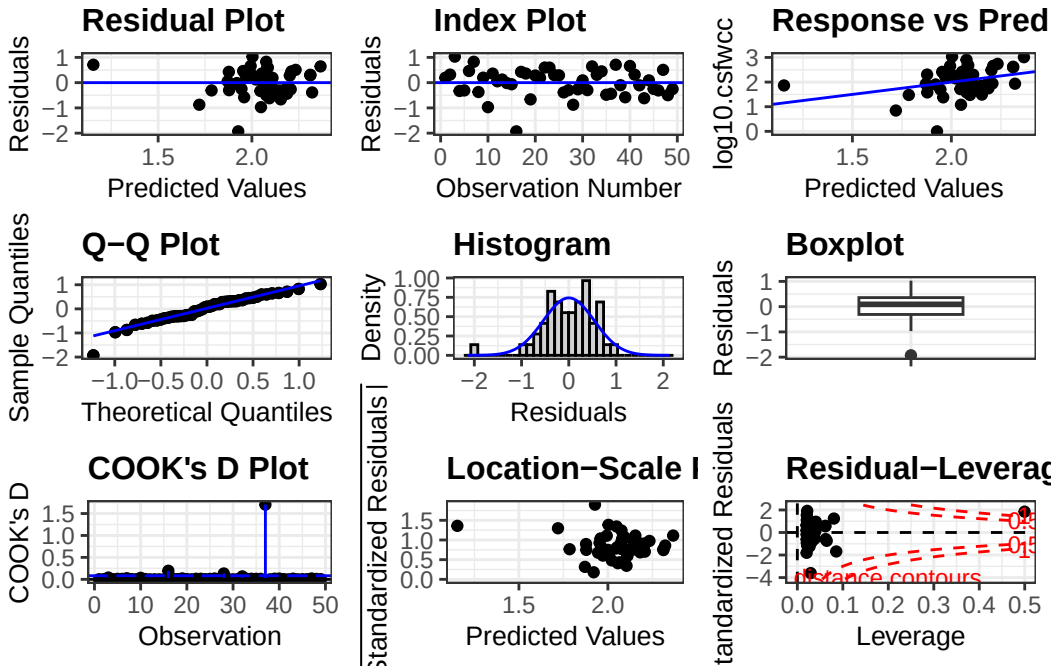
```
# As a variation, we add a boxplot instead of a histogram
p <- ggplot(cm.hivneg, aes(log10.bldwcc,log10.csfwcc)) + geom_point()
ggMarginal( p, type = "boxplot")
```




```
p + geom_smooth(method="lm")
```




```
# We use the resid\_panel function from the ggResidpanel package, and now plot all res
resid_panel(fit.log,"all")
```



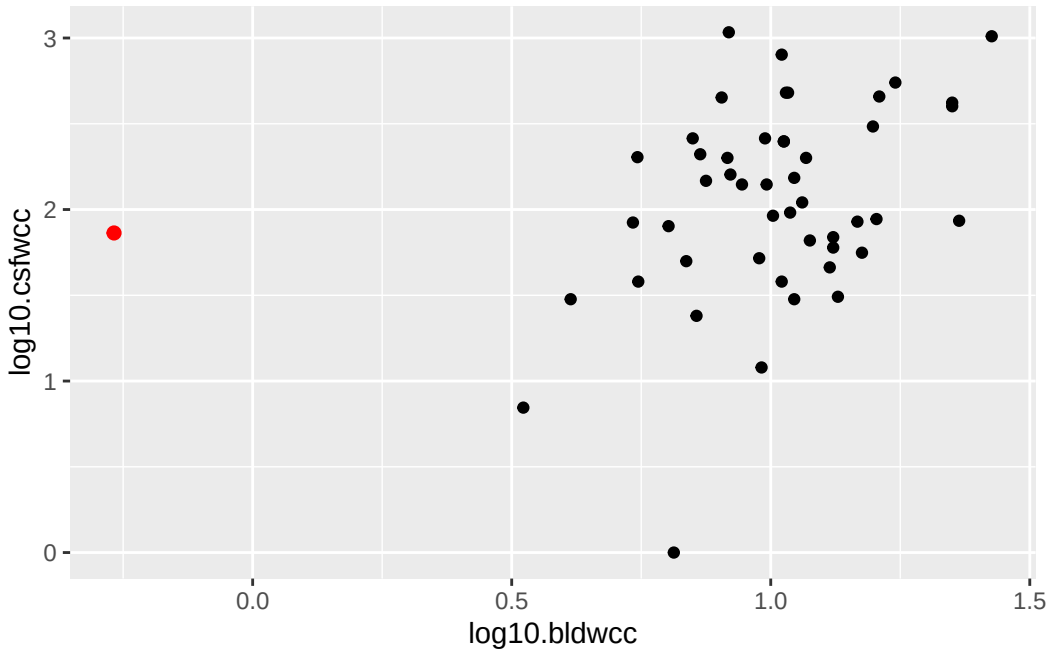

```
# The extreme individual is the 37th observation
```

e) The "Residuals vs Leverage" plot suggests that there's one individual that may have a large impact on the parameter estimates. Identify this point and perform steps b) and c) again for the log-transformed data with that observation removed. Comments? Do you see any other individual that may not follow the model assumptions?

```
cm.hivneg["149",]
```

```
      code hiv diagnosis group    groupLong age sex bldwcc bldneut bldlym
149 BMD901   0        CM     4 HIV neg - CM  49   2   0.54    0.43   0.05
    csfwcc csfneut csflym log10.bldwcc log10.csfwcc
149     73   64.24    7.3   -0.2676062     1.863323
```

```
# The individual in row number 149 is extreme
# Note that the subset function keeps the row numbers as in the original data set cmTb
# Hence, that individual is not the 149th row in cm.hivneg
# We can visualize and show its values via
ggplot(cm.hivneg, aes(log10.bldwcc, log10.csfwcc)) +
  geom_point() +
  geom_point(data = cm.hivneg["149",], colour = "red", size = 2)
```




```
# In fact, it is the only person with a negative value for log10.bldwcc, hence we can
subset(cm.hivneg, log10.bldwcc < 0)
```

```
      code hiv diagnosis group    groupLong age sex bldwcc bldneut bldlym
149 BMD901   0        CM     4 HIV neg - CM  49   2   0.54    0.43   0.05
    csfwcc csfneut csflym log10.bldwcc log10.csfwcc
149     73   64.24    7.3   -0.2676062     1.863323
```

```
# Refit without that person
fit.log.del <- lm(log10.csfwcc ~ log10.bldwcc, data = cm.hivneg, subset = row.names(cm
# Easier is
# fit.log.del <- lm(log10.csfwcc ~ log10.bldwcc, data = cm.hivneg, subset = log10.bldw
# Or observing that it is individual 37
# fit.log.del <- lm(log10.csfwcc ~ log10.bldwcc, data = cm.hivneg[-37,])
summary(fit.log.del)
```

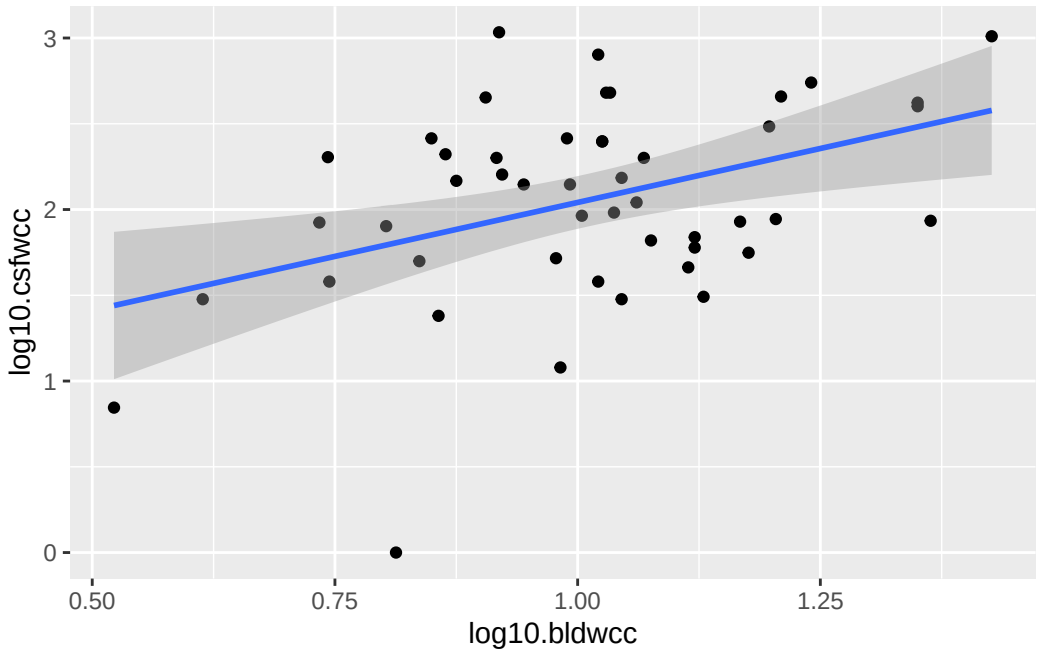
```
Call:
lm(formula = log10.csfwcc ~ log10.bldwcc, data = cm.hivneg, subset = row.names(cm.hivn
    "149")
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.8059 -0.3538  0.1124  0.3567  1.0940

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.7829     0.4200   1.864  0.06868 .
log10.bldwcc   1.2583     0.4090   3.076  0.00352 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5284 on 46 degrees of freedom
Multiple R-squared:  0.1706,    Adjusted R-squared:  0.1526
F-statistic: 9.465 on 1 and 46 DF,  p-value: 0.003522
```

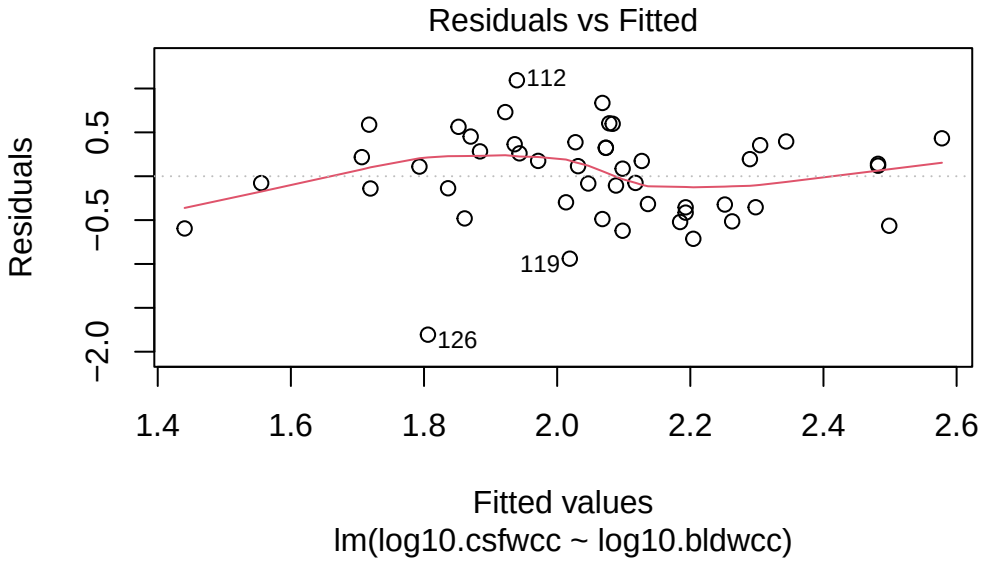
```r
confint(fit.log.del)
```

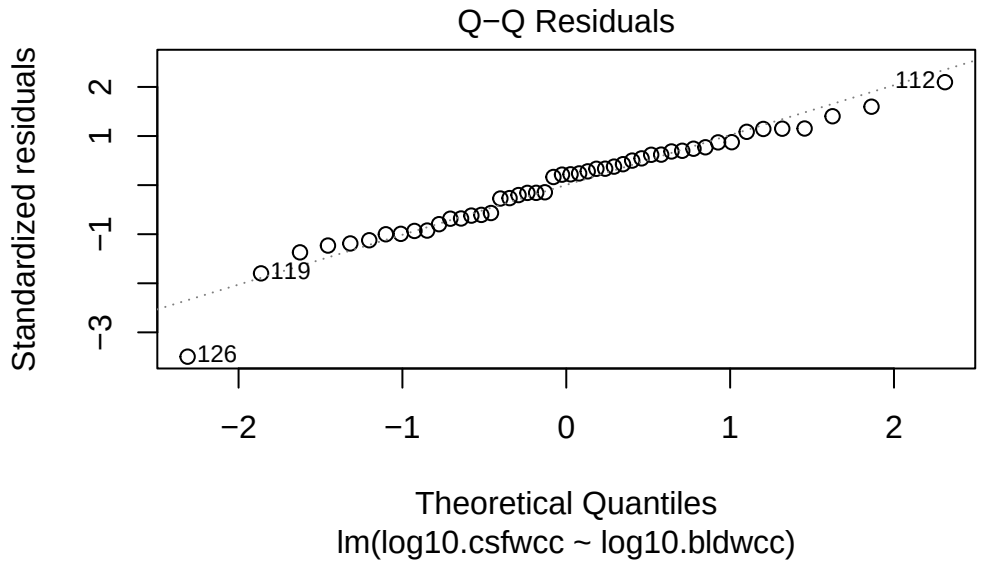
```
                  2.5 %   97.5 %
(Intercept)  -0.06242973 1.628321
log10.bldwcc  0.43502510 2.081659
```

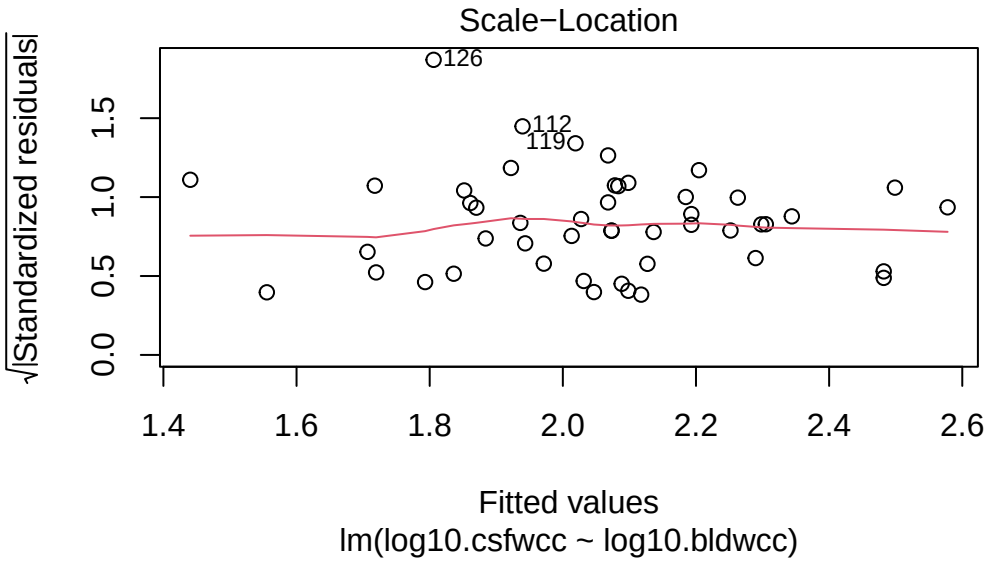
```r
ggplot(subset(cm.hivneg, log10.bldwcc > 0), aes(log10.bldwcc, log10.csfwcc)) +
  geom_point() +
  geom_smooth(method = "lm")
```

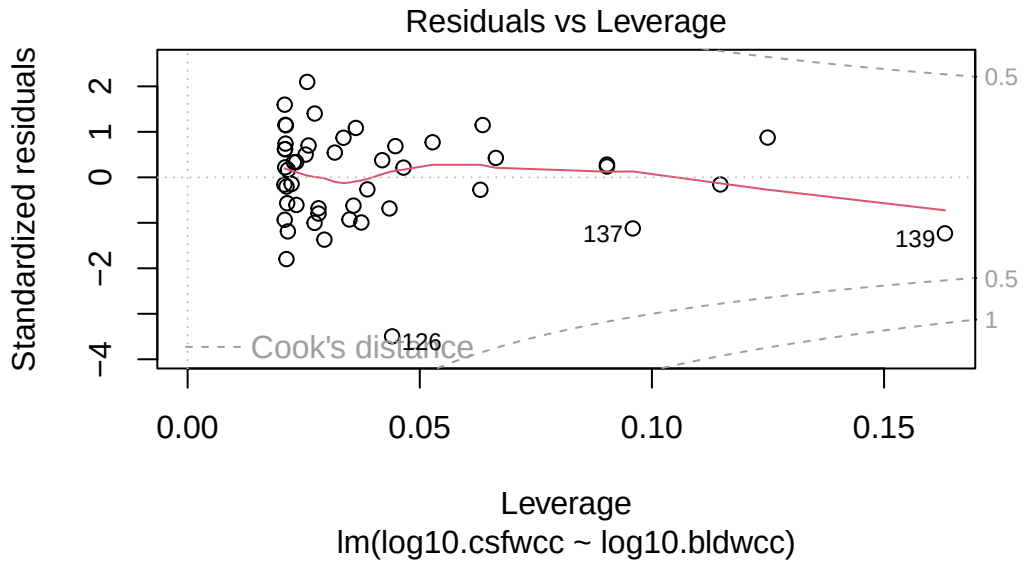



```
plot(fit.log.del)
```




Q−Q Residuals
Standardized residuals
112
119
126
−3
−1
1
2
−2
−1
0
1
2
Theoretical Quantiles
lm(log10.csfwcc ~ log10.bldwcc)


Scale−Location
√|Standardized residuals|
126
112
119
0.0
0.5
1.0
1.5
1.4
1.6
1.8
2.0
2.2
2.4
2.6
Fitted values
lm(log10.csfwcc ~ log10.bldwcc)




```
# Now individual with row number 126 may be problematic with respect to the QQ plot an
fit.log.del2 <- lm(log10.csfwcc ~ log10.bldwcc, data = cm.hivneg, subset = !row.names(
summary(fit.log.del2)
```

```
Call:
lm(formula = log10.csfwcc ~ log10.bldwcc, data = cm.hivneg, subset = !row.names(cm.hiv
    c("126", "149"))

Residuals:
     Min       1Q   Median       3Q      Max
-0.98527 -0.34672  0.07149  0.32415  1.03442

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0472     0.3697   2.832  0.00689 **
log10.bldwcc   1.0356     0.3587   2.887  0.00595 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4578 on 45 degrees of freedom
Multiple R-squared:  0.1563,    Adjusted R-squared:  0.1376
F-statistic: 8.336 on 1 and 45 DF,  p-value: 0.005952
```

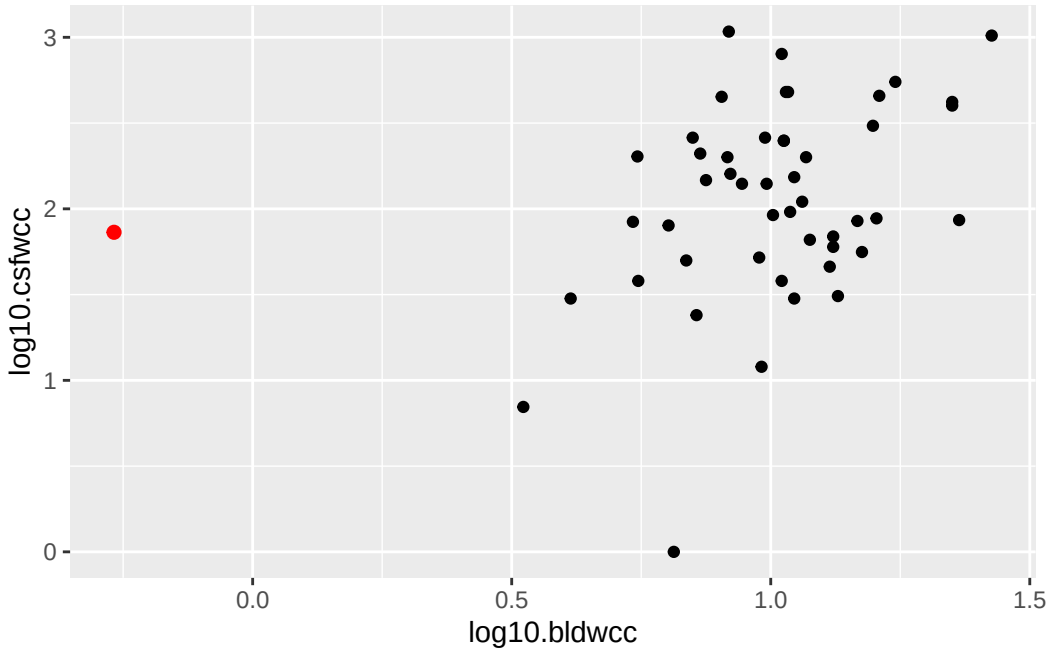
```
confint(fit.log.del2)
```

```
                2.5 %   97.5 %
(Intercept)  0.3025658 1.791921
log10.bldwcc 0.3131804 1.757949
```

Answer e: 1. Especially the individual in row number 149 is extreme. 2. Fit model again without that individual. The estimates change quite a lot. The p-value has gone down again. Diagnostic plot looks better. 3. Now individual with row number 126 may be problematic with respect the QQ plot and Residuals vs Fitted. Again, things change quite a bit. However, we should be prudent in removing individuals. It is partly subjective when to stop. Just report your decisions in your analysis, and keep in mind that estimates and p-values can be quite sensitive to such decisions.

Alternative solution:

```
# the individual in row 37 is extreme
# we can visualize and show its values via
ggplot(cm.hivneg, aes(log10.bldwcc,log10.csfwcc)) + geom_point() +
  geom_point(data=cm.hivneg[37,], colour="red", size=2)
```



```
subset(cm.hivneg, log10.bldwcc<0 )
```

```
      code hiv diagnosis group    groupLong age sex bldwcc bldneut bldlym
149 BMD901   0        CM     4 HIV neg - CM  49   2   0.54    0.43   0.05
    csfwcc csfneut csflym log10.bldwcc log10.csfwcc
149     73   64.24    7.3   -0.2676062     1.863323
```

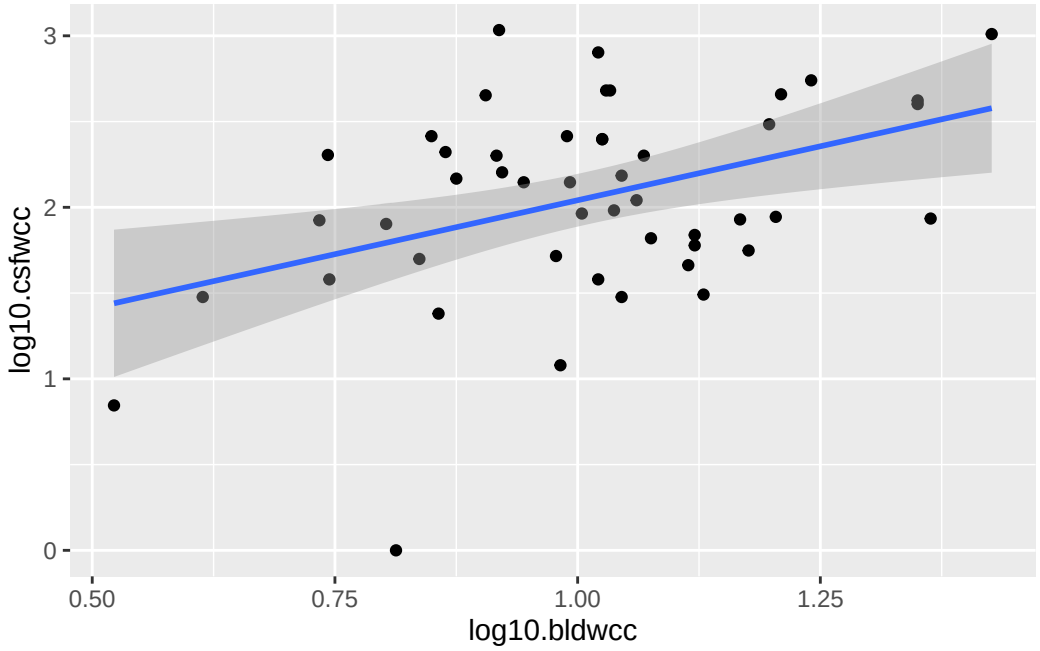
```
# -> refit without observations  37
fit.log.del.alt <- lm(log10.csfwcc~log10.bldwcc, data=cm.hivneg[-37,])
summary(fit.log.del.alt)
```

```
Call:
lm(formula = log10.csfwcc ~ log10.bldwcc, data = cm.hivneg[-37,
    ])

Residuals:
    Min      1Q  Median      3Q     Max
-1.8059 -0.3538  0.1124  0.3567  1.0940

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.7829     0.4200   1.864  0.06868 .
log10.bldwcc   1.2583     0.4090   3.076  0.00352 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5284 on 46 degrees of freedom
Multiple R-squared:  0.1706,    Adjusted R-squared:  0.1526
F-statistic: 9.465 on 1 and 46 DF,  p-value: 0.003522
```

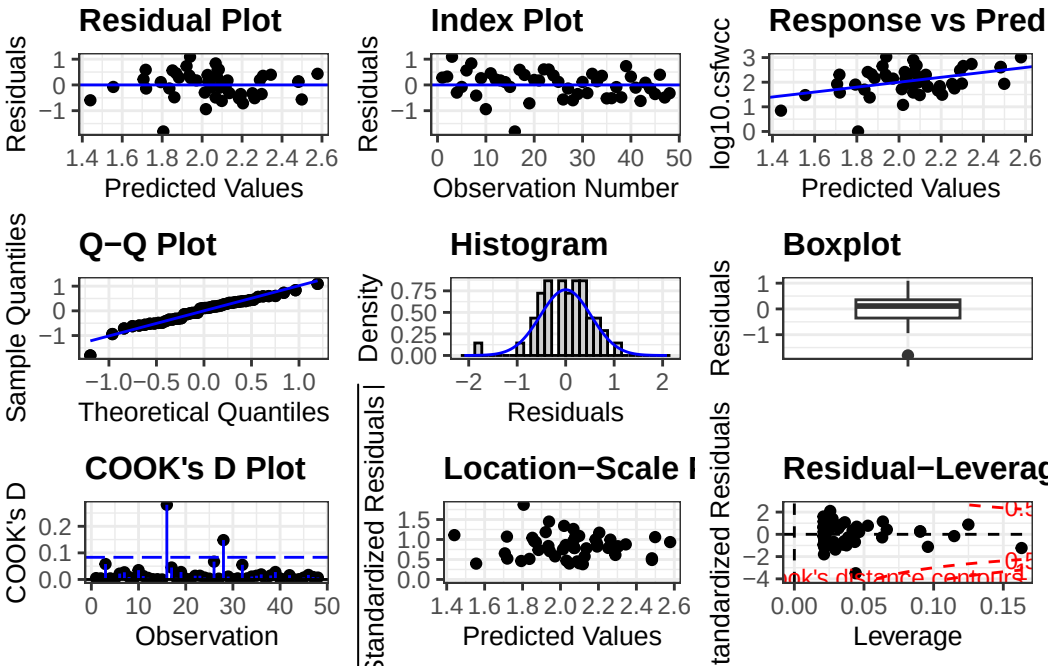
```
confint(fit.log.del.alt)
```

```
                  2.5 %   97.5 %
(Intercept)  -0.06242973 1.628321
log10.bldwcc  0.43502510 2.081659
```

```
ggplot(cm.hivneg[-37,], aes(log10.bldwcc,log10.csfwcc)) + geom_point() + geom_smooth(m
```




```
resid_panel(fit.log.del.alt, plots="all")
```



```r
# Now individual 16 seems to be problematic
fit.log.del2.alt <- lm(log10.csfwcc~log10.bldwcc, data=cm.hivneg[-c(16,37),])
summary(fit.log.del2.alt)
```

```
Call:
lm(formula = log10.csfwcc ~ log10.bldwcc, data = cm.hivneg[-c(16,
    37), ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.98527 -0.34672  0.07149  0.32415  1.03442

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0472     0.3697   2.832  0.00689 **
log10.bldwcc   1.0356     0.3587   2.887  0.00595 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4578 on 45 degrees of freedom
Multiple R-squared:  0.1563,    Adjusted R-squared:  0.1376
F-statistic: 8.336 on 1 and 45 DF,  p-value: 0.005952
```

```r
confint(fit.log.del2.alt)
```

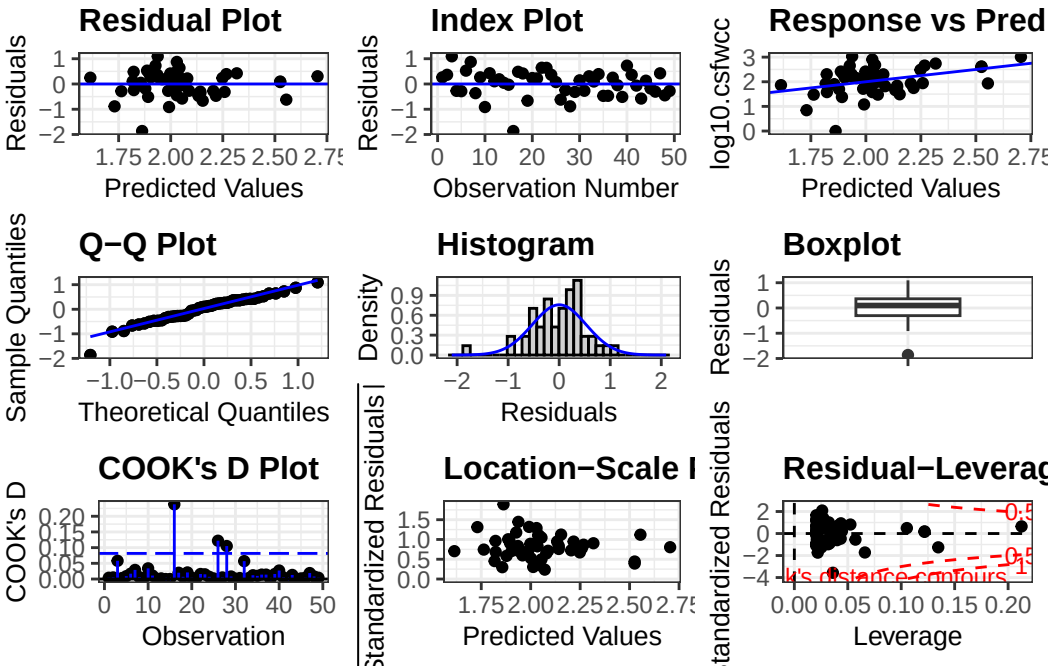
```
                2.5 %   97.5 %
(Intercept)  0.3025658 1.791921
log10.bldwcc 0.3131804 1.757949
```

```r
# we can also decide to transform only the dependent variable, csfwcc (it was the extr
# blood cell count that led to the outlier on the log scale)
# remember that we don't need to assume approximate normality for the covariable, here
fit.log.csf <- lm(log10.csfwcc~bldwcc, data=cm.hivneg)
summary(fit.log.csf)
```

```
Call:
```

```
lm(formula = log10.csfwcc ~ bldwcc, data = cm.hivneg)

Residuals:
     Min       1Q   Median       3Q      Max
-1.86267 -0.30318  0.09761  0.36432  1.09570

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.59164    0.17975   8.855 1.39e-11 ***
bldwcc       0.04170    0.01483   2.811  0.00718 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5317 on 47 degrees of freedom
Multiple R-squared:  0.1439,    Adjusted R-squared:  0.1257
F-statistic: 7.903 on 1 and 47 DF,  p-value: 0.007176
```

```
resid_panel(fit.log.csf, plots="all")
```



f) What CSF white cell count does the model from e) predict for a patient with a white cell count in blood of 10x10^3/mm^3, i.e., with log10.bldwcc = 1? Calculate a 95% prediction interval for log10.csfwcc in a patient with log10.bldwcc = 1.

```r
# Prediction of csfwcc for a patient with lo10bldwcc=1
predict(fit.log.del,newdata=data.frame(log10.bldwcc=1),interval="prediction")
```

```
       fit      lwr      upr
1 2.041288 0.966643 3.115932
```

```r
predict(fit.log.del,newdata=data.frame(log10.bldwcc=1),interval="confidence")
```

```
       fit      lwr      upr
1 2.041288 1.887563 2.195012
```

```r
# The prediction interval is much wider than the confidence interval.

# Slightly easier is to use the following code from the ggeffects package. Note that r
# exactly the same because ggeffects uses the normal distribution instead of the t-dis
# You may have to install that package first
library(ggeffects)
ggpredict(fit.log.del)
```

```
$log10.bldwcc
# Predicted values of log10.csfwcc

log10.bldwcc | Predicted |     95% CI
-------------------------------------
        0.50 |      1.41 | 0.97, 1.86
        0.60 |      1.54 | 1.17, 1.91
        0.70 |      1.66 | 1.37, 1.96
        0.80 |      1.79 | 1.56, 2.02
        1.00 |      2.04 | 1.89, 2.20
        1.10 |      2.17 | 2.00, 2.34
        1.20 |      2.29 | 2.07, 2.51
        1.40 |      2.54 | 2.19, 2.90


attr(,"class")
[1] "ggalleffects" "list"
attr(,"model.name")
[1] "fit.log.del"
```

```
ggpredict(fit.log.del, interval="prediction")
```

```
$log10.bldwcc
# Predicted values of log10.csfwcc

log10.bldwcc | Predicted |     95% CI
-------------------------------------
        0.50 |      1.41 | 0.26, 2.57
        0.60 |      1.54 | 0.41, 2.66
        0.70 |      1.66 | 0.56, 2.77
        0.80 |      1.79 | 0.70, 2.88
        1.00 |      2.04 | 0.97, 3.12
        1.10 |      2.17 | 1.09, 3.24
        1.20 |      2.29 | 1.21, 3.38
        1.40 |      2.54 | 1.42, 3.67


attr(,"class")
[1] "ggalleffects" "list"
attr(,"model.name")
[1] "fit.log.del"
```