Lab 4 Part 2 - Solution

Isabel Sassoon

October 2020

```
library(ggplot2)
```

(1) Read in the data

```
birth<-read.csv("data/birthweight.csv", sep = ";")
```

(2) Explore the data numerically and graphically

```
summary(birth)
```

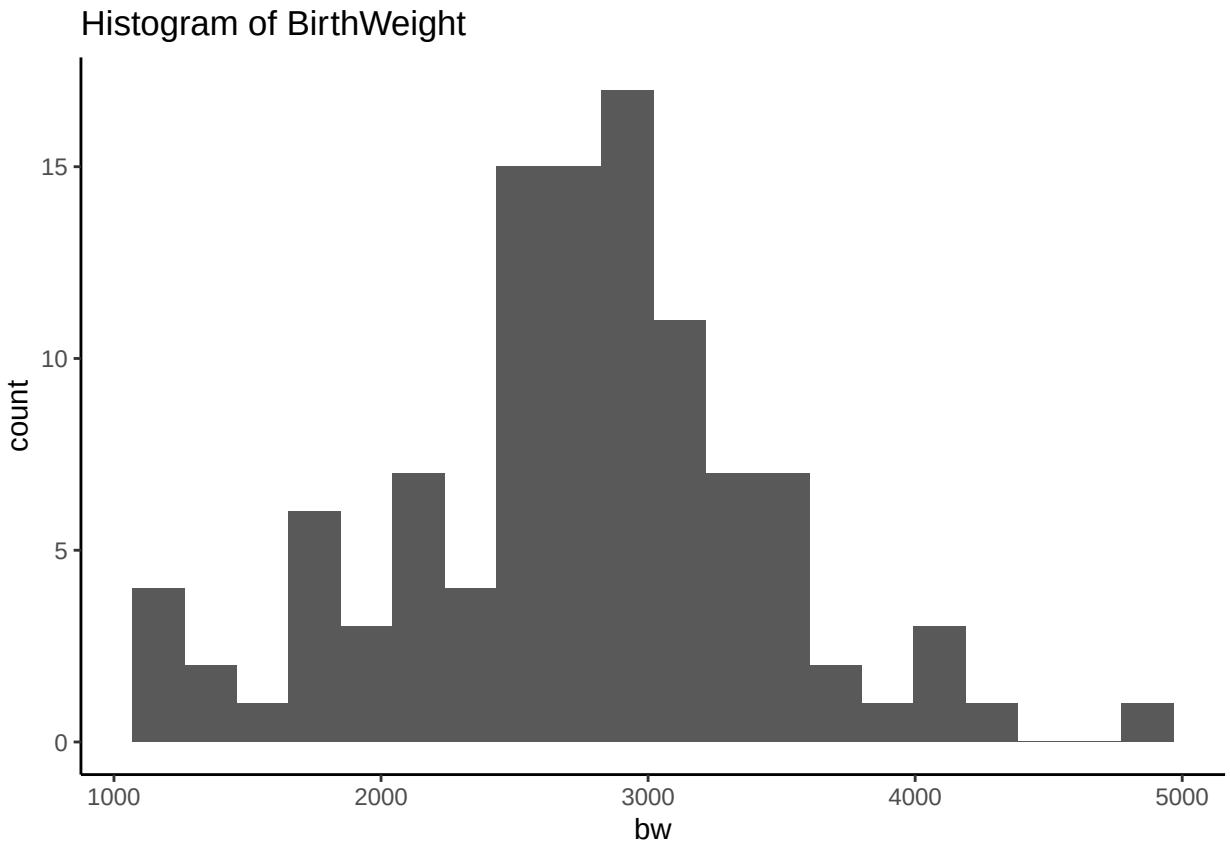
```
##        bw            bpd               ad             idnr
##  Min.   :1150   Min.   : 64.00   Min.   : 71.0   Min.   :  1.0
##  1st Qu.:2400   1st Qu.: 88.00   1st Qu.: 96.0   1st Qu.: 27.5
##  Median :2800   Median : 91.00   Median :103.0   Median : 54.0
##  Mean   :2739   Mean   : 89.48   Mean   :101.7   Mean   : 54.0
##  3rd Qu.:3125   3rd Qu.: 93.00   3rd Qu.:108.0   3rd Qu.: 80.5
##  Max.   :4850   Max.   :100.00   Max.   :133.0   Max.   :107.0
```

This data has 4 attributes as expected. The numerical summaries don't highlight any issues related to the columns, such as missing values or extreme/skewed values.

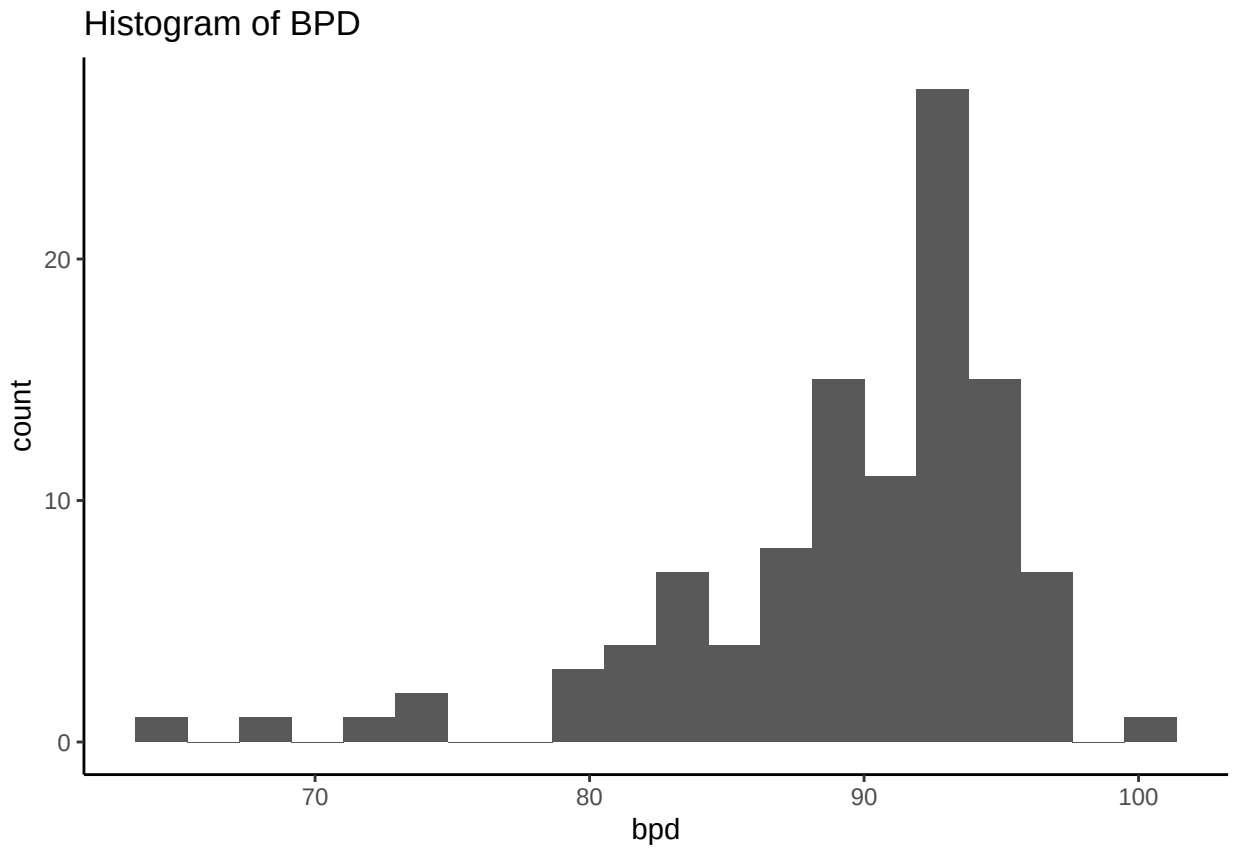
Visualise the data

The next step is to visualise this data (explore it graphically). It is possible to use a histogram to look at each variable's distribution.

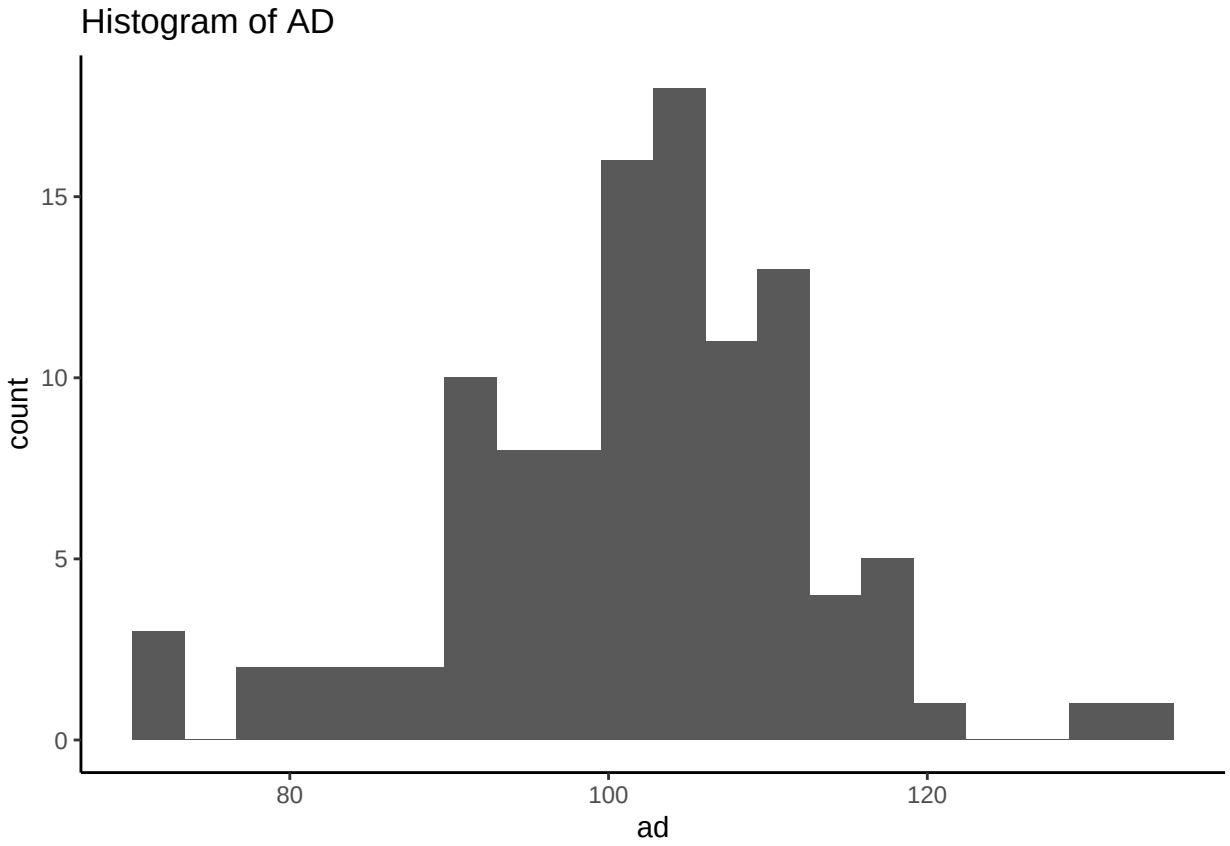
```
ggplot(data=birth, aes(x=bw)) + geom_histogram(bins = 20) + theme_classic() +ggtitle("Histogram of Birtl
```




```
ggplot(data=birth, aes(x=bpd)) + geom_histogram(bins = 20) + theme_classic() +ggtitle("Histogram of BPD
```




```
ggplot(data=birth, aes(x=ad)) + geom_histogram(bins = 20) + theme_classic() +ggtitle("Histogram of AD")
```

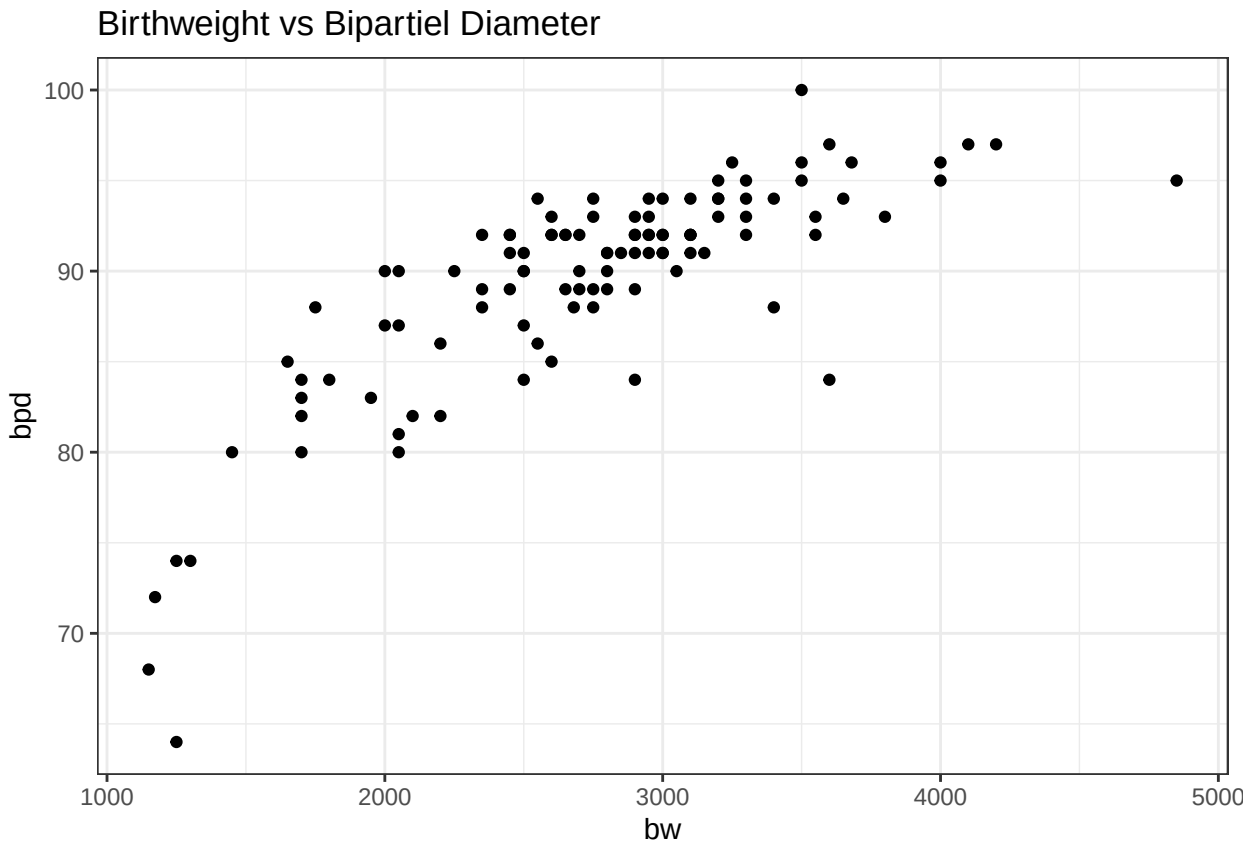


The variable idnr is just an identifier, so not useful for the analysis.

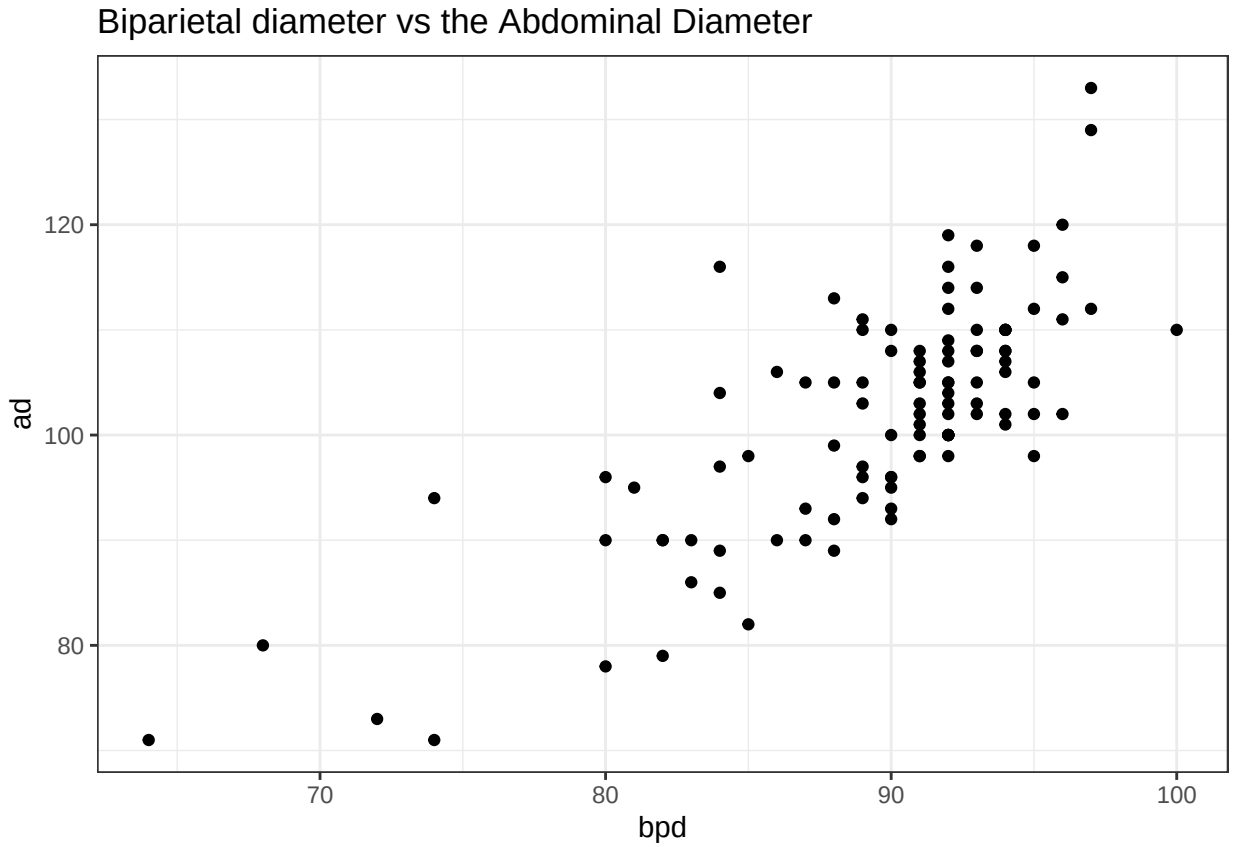
The three histograms show us that bw and ad seem to have a symmetric normal-like distribution, whereas bpd is skewed to the right. There are no visible extreme outliers to consider at the moment. (In real world situations this would be a good output to share with the clinicians for them to confirm that the data looks as expected)

Now we can move to plotting the relationship between the three columns:

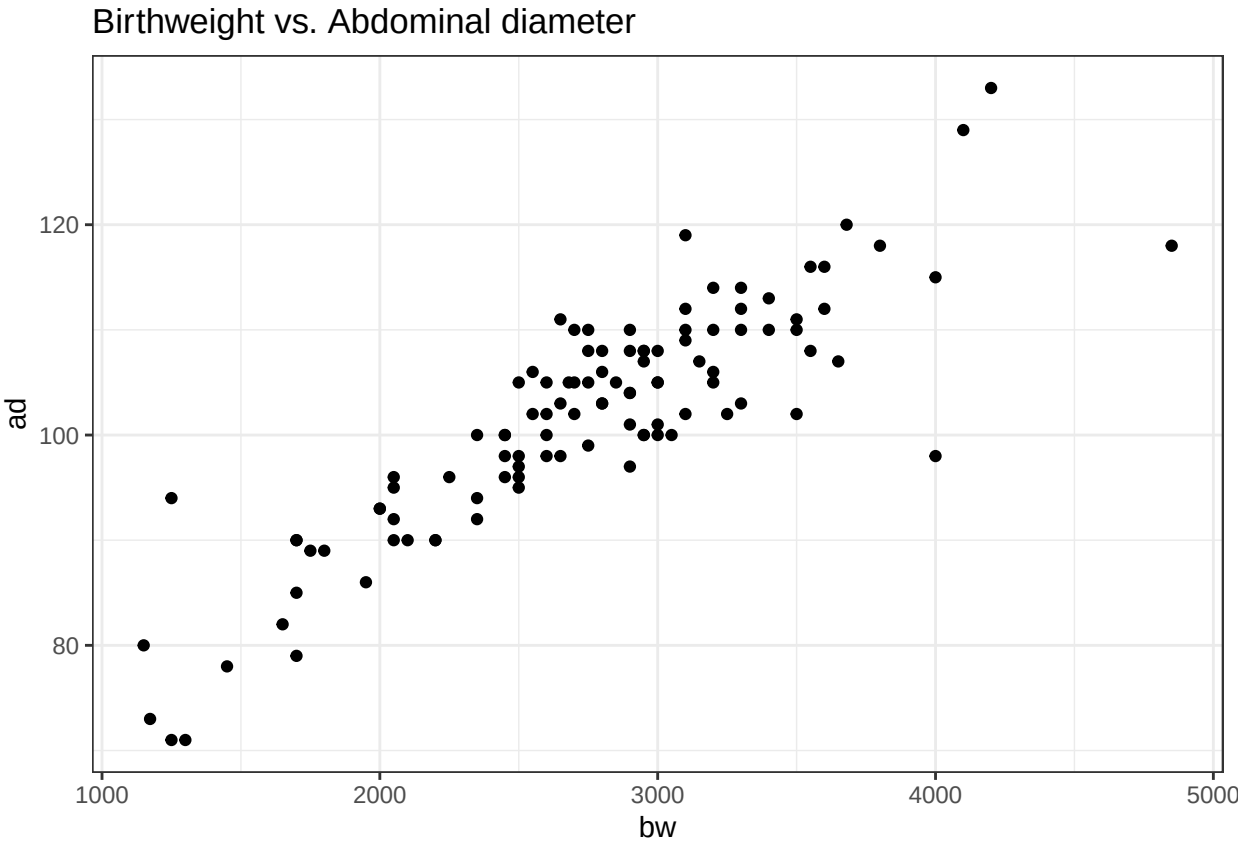
```
ggplot(birth, aes(x=bw, y=bpd))+ geom_point() + theme_bw() + ggtitle("Birthweight vs Bipartiel Diameter
```




```
ggplot(birth, aes(x=bpd, y=ad))+ geom_point() + theme_bw() + ggtitle("Biparietal diameter vs the Abdomi
```




```
ggplot(birth, aes(x=bw, y=ad))+ geom_point() + theme_bw() + ggtitle("Birthweight vs. Abdominal diameter
```



There seems to be potenital for linear relations betwen the attributes.

Correlation

In order to assess the strength of the linear relation between BW and each of the other variables in turn we can use the correlation.

```
cor(birth$bw, birth$bpd)
```

```
## [1] 0.7980023
```

```
cor(birth$bw, birth$ad)
```

```
## [1] 0.8730657
```

and we can also test if these correlations are significant

```
cor.test(birth$bw, birth$bpd)
```

```
##
##  Pearson's product-moment correlation
##
## data:  birth$bw and birth$bpd
## t = 13.568, df = 105, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7167345 0.8578848
## sample estimates:
##       cor
```

```
## 0.7980023
```

```
cor.test(birth$bw, birth$ad)
```

```
##
##  Pearson's product-moment correlation
##
## data:  birth$bw and birth$ad
## t = 18.347, df = 105, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8189566 0.9117873
## sample estimates:
##       cor
## 0.8730657
```

Both correlations are significant, but the one betwen BW and AD appears stronger.

Build two regression models

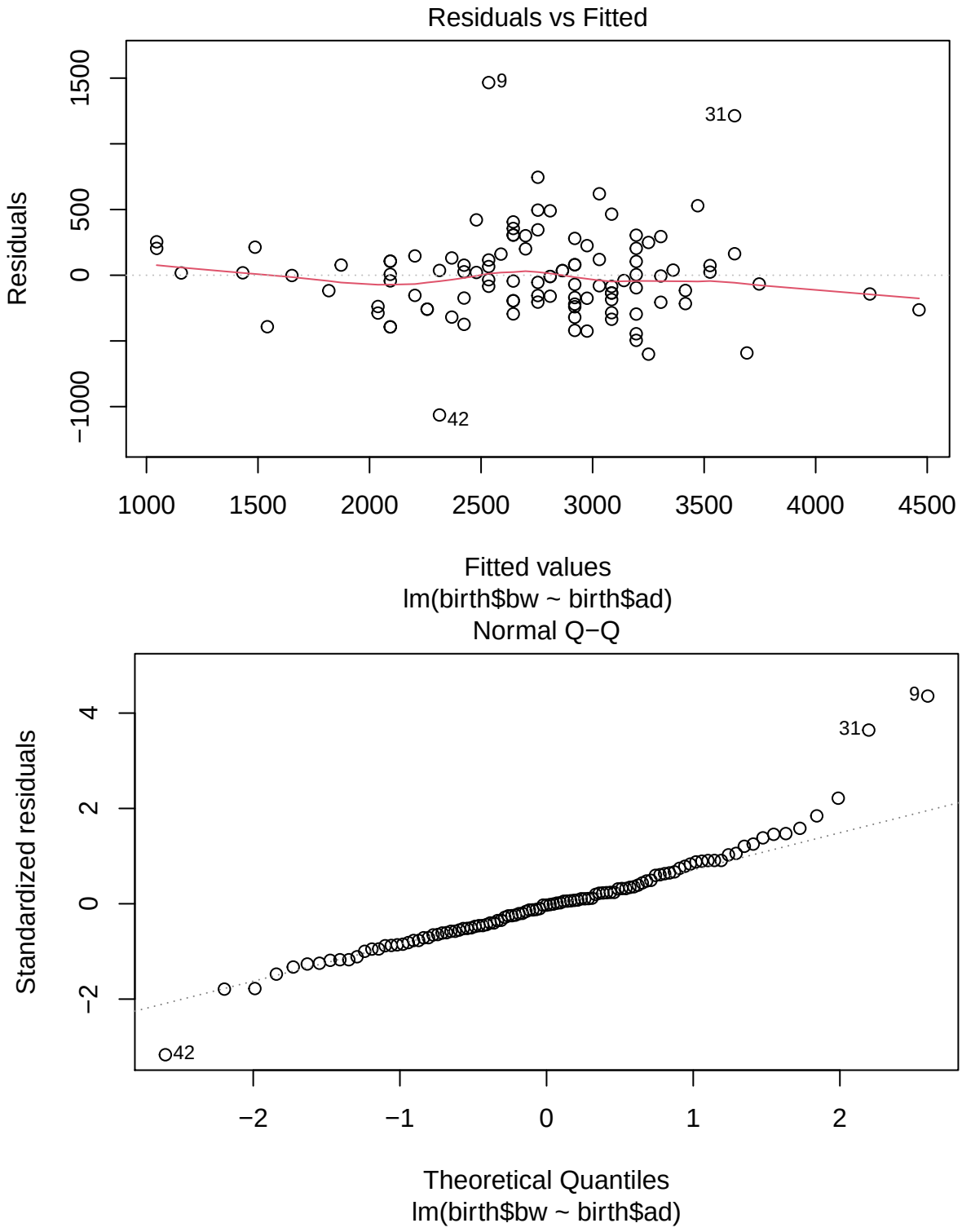
The birth weight (BW) is our dependent variable as this is what we would want to predict based on the other measurements.

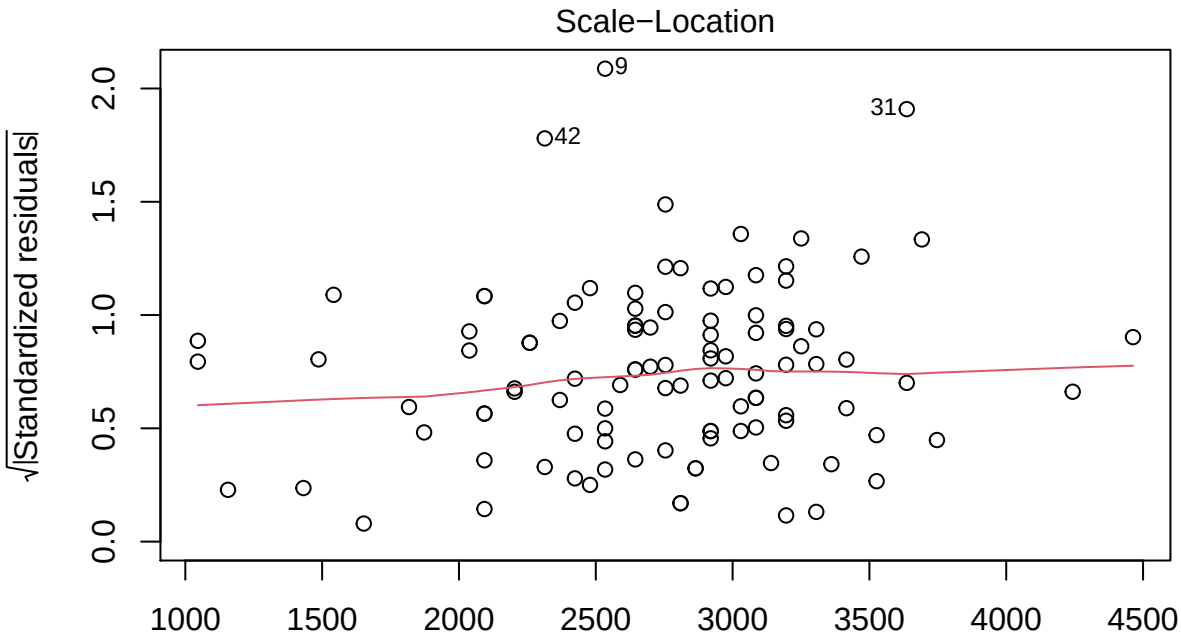
```
ad.lm<-lm(birth$bw~birth$ad)
summary(ad.lm)
```

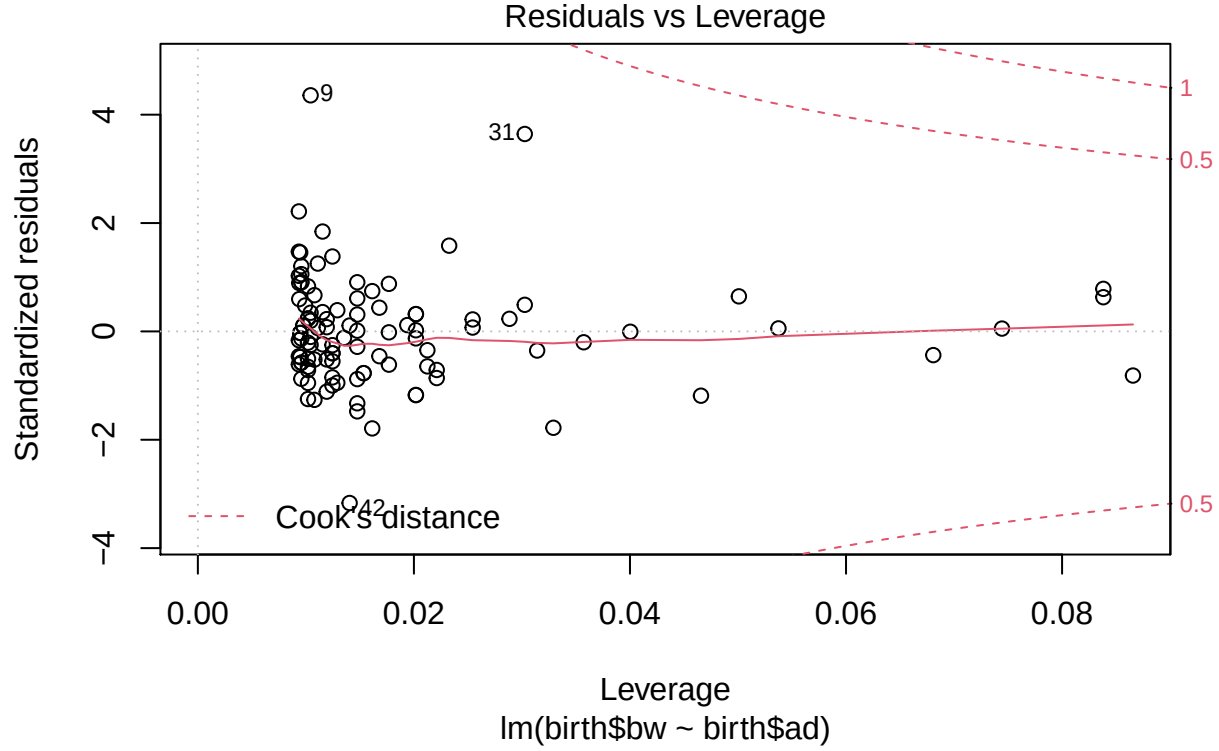
```
##
## Call:
## lm(formula = birth$bw ~ birth$ad)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1063.57  -199.43    -9.67   153.75  1465.94
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2867.916    307.346  -9.331 1.94e-15 ***
## birth$ad       55.122      3.004  18.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 338.2 on 105 degrees of freedom
## Multiple R-squared:  0.7622, Adjusted R-squared:   0.76
## F-statistic: 336.6 on 1 and 105 DF,  p-value: < 2.2e-16
```

and the graphical diagnostics

```
plot(ad.lm)
```


Residuals vs Fitted
Residuals
1500
500
0
−1000
9
31
42
1000
1500
2000
2500
3000
3500
4000
4500
Fitted values
lm(birth$bw ~ birth$ad)
Normal Q−Q
Standardized residuals
4
2
0
−2
9
31
42
−2
−1
0
1
2
Theoretical Quantiles
lm(birth$bw ~ birth$ad)





This model can be written as:

$$bw = -2867.92 + 55.12 \times ad$$

There is a positive relation between bw and ad. An increase of 1 in ad results in an estimated increase of 55.12 in the bw.

This first model has significant coefficients, a significant difference in the SSR SSE ratio (see F-Test) and a high r^2. The diagnostic plots do not point to major issues with the model, but there are some possible values

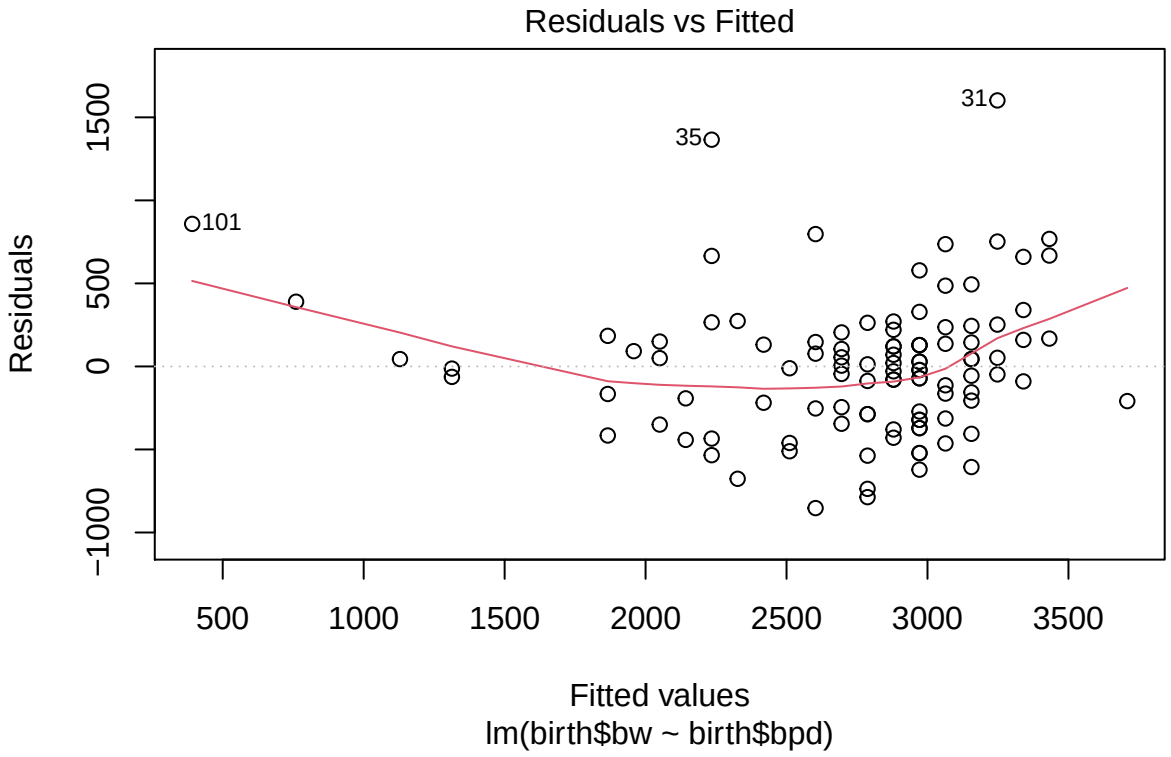
that look like outliers (row 9).

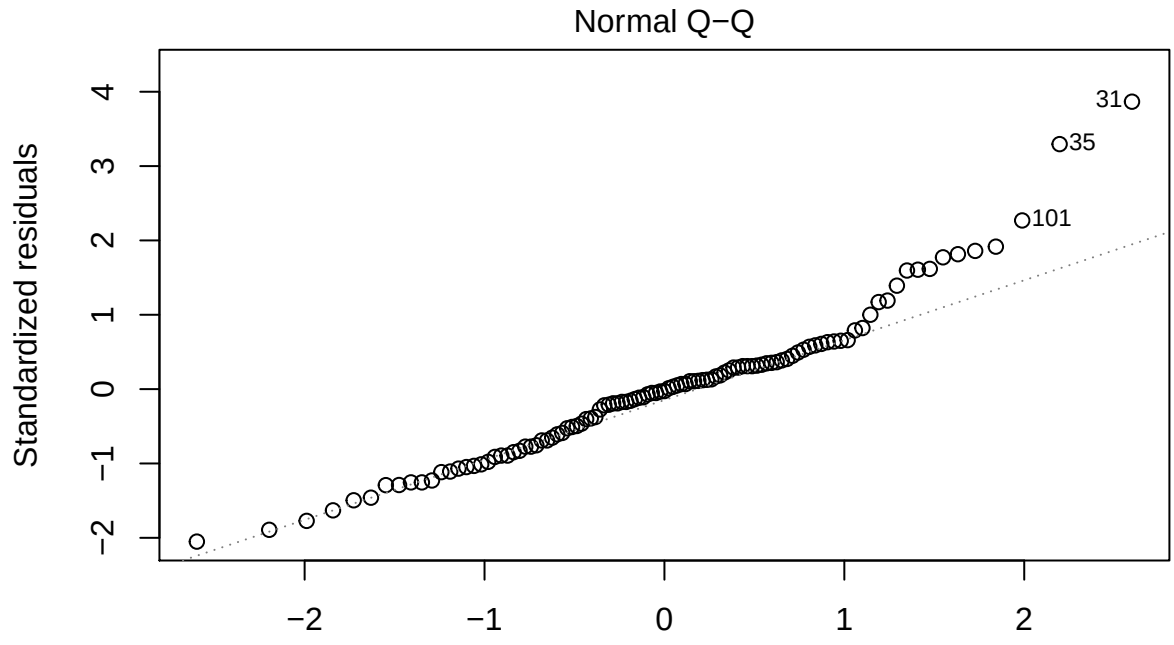
Lets build the second model with bpd as the explanatory variable:

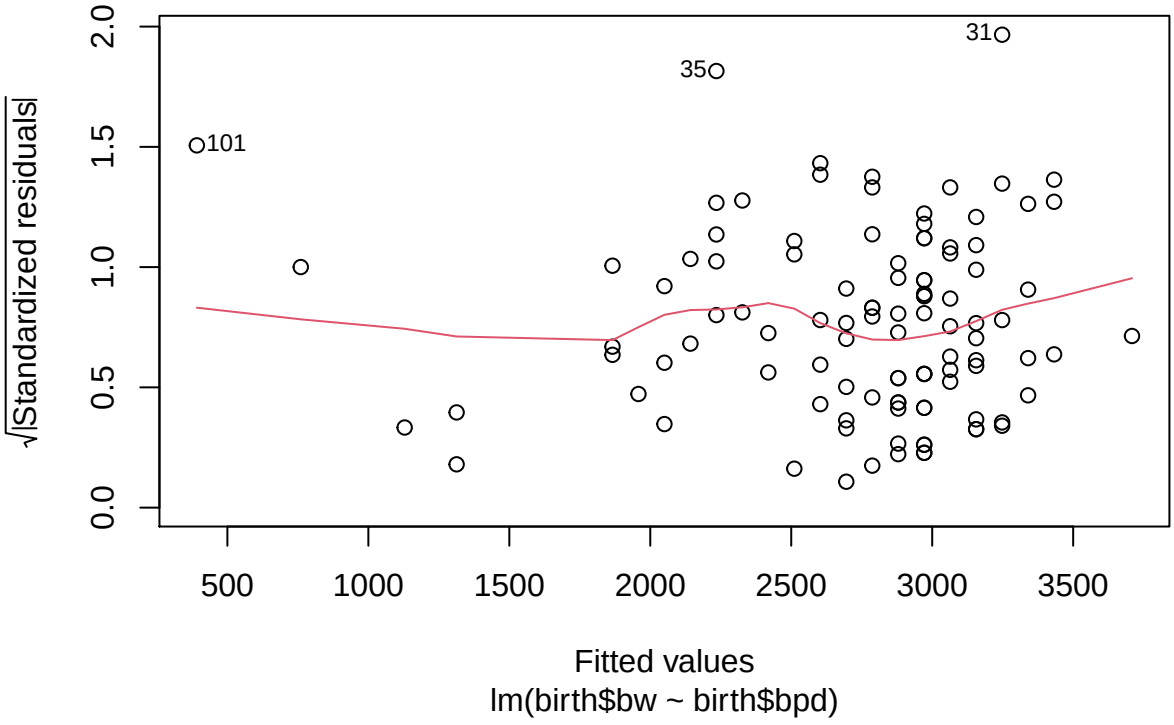
```
bpd.lm<-lm(birth$bw~birth$bpd)
summary(bpd.lm)
```

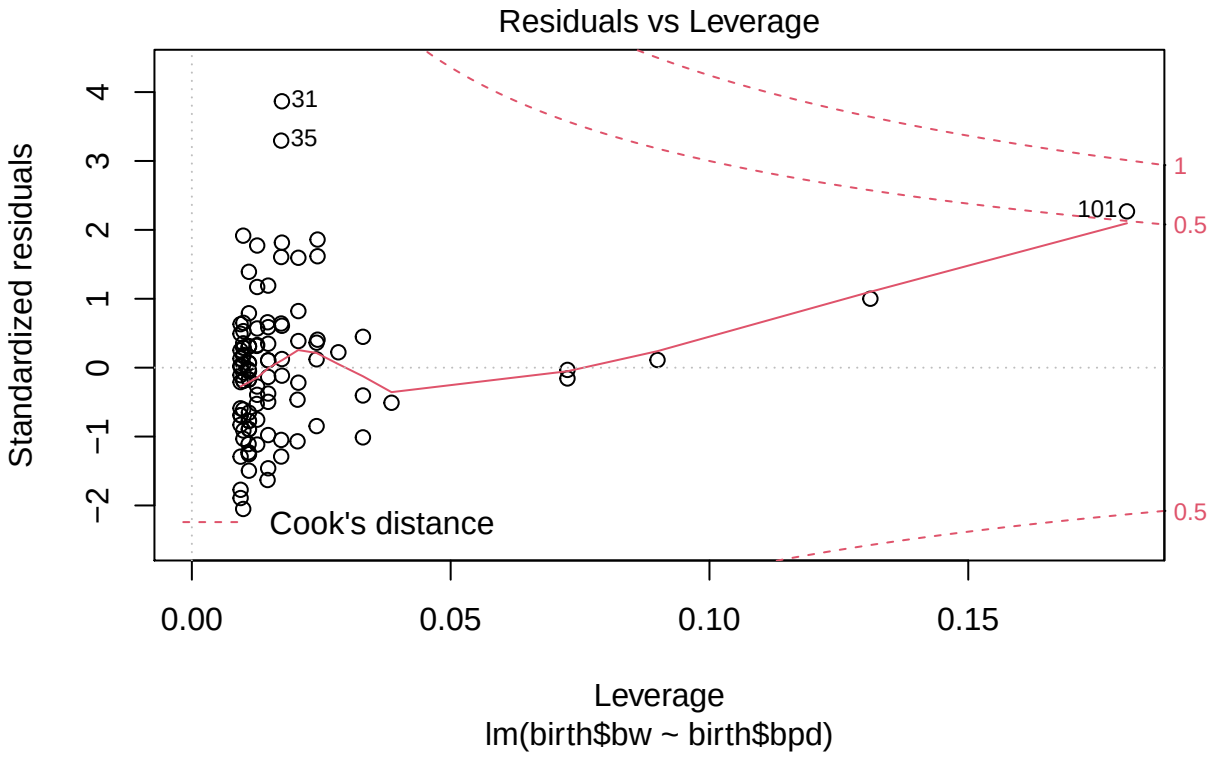
```
##
## Call:
## lm(formula = birth$bw ~ birth$bpd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -853.03 -287.32  -10.89  163.76 1601.98
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5505.406    608.965  -9.041 8.67e-15 ***
## birth$bpd      92.141      6.791  13.568  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 418 on 105 degrees of freedom
## Multiple R-squared:  0.6368, Adjusted R-squared:  0.6333
## F-statistic: 184.1 on 1 and 105 DF,  p-value: < 2.2e-16
```

```
plot(bpd.lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(birth$bw ~ birth$bpd)
31
35
101


Scale−Location
√|Standardized residuals|
Fitted values
lm(birth$bw ~ birth$bpd)
31
35
101



This model can be written as:

$$bw = -5505.41 + 92.14 \times bpd$$

We can see that there is a positive relation between the bpd and the bw. An increase in 1 in bpd results in an estimated bw increase of 92.14.

This model also has significant coefficients, a significant difference in the SSR SSE ratio (see F-Test) and a high r^2. The diagnostic plots do not point to major issues with the model, but the first plot could be a clue that there is a pattern in the residuals. The r^2 is also lower than the previous model.

(5) Comparing the two models. Which is better at predicting and why?

The R^2 is higher for the model that uses AD. But both are good models.The diagnostic plots are acceptable in both cases, however in our first model there is potential to investigate outliers and in the second model there is a pattern in the residuals that may warrant a transformation. In this case the model that uses AD would be preferred for predicting, as it has a higher r^2.

(6) Try some transformations

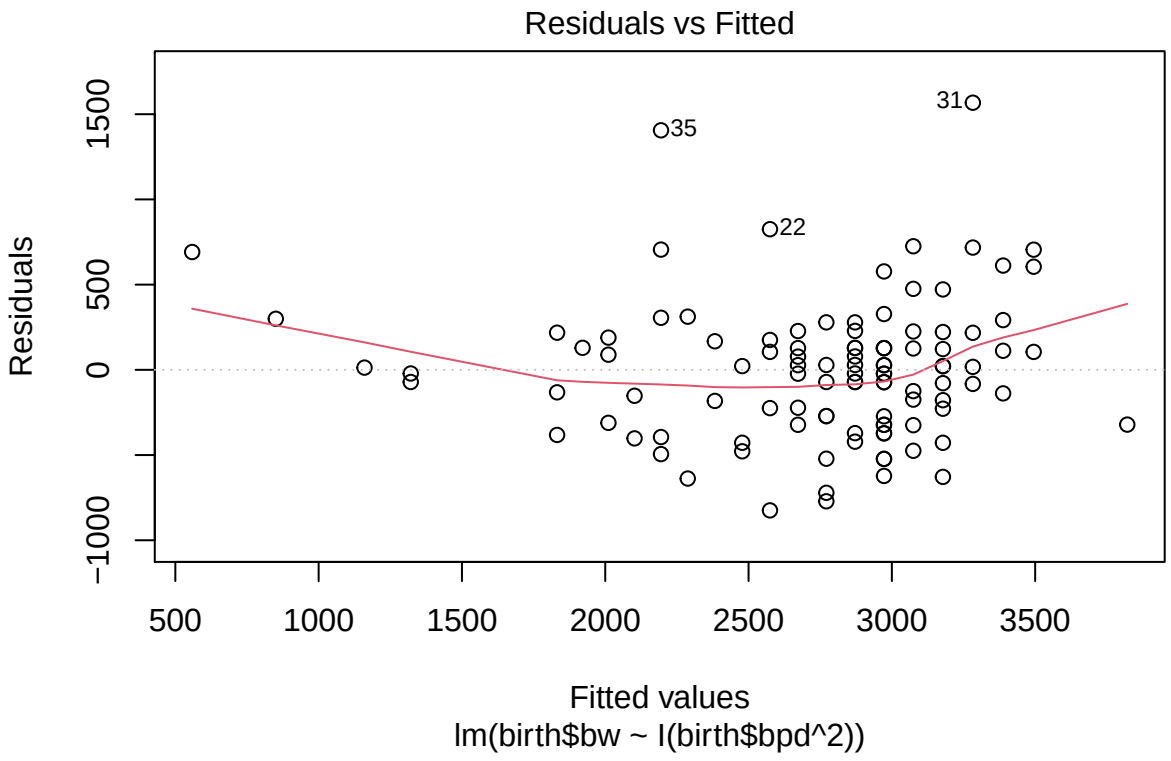
There is a pattern in the model that uses BPD, see the residuals (first diagnostic plot), perhaps a polynomial model would be an effective option?

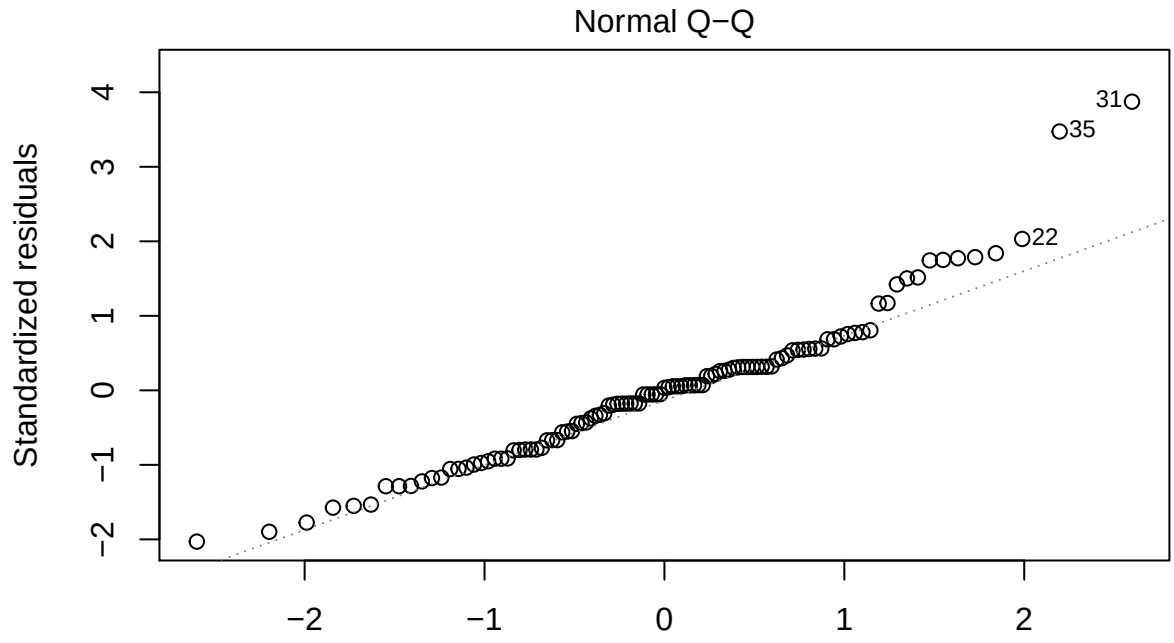
Note: I() is a way of computing the transformation as part of the model definition. It means we dont actually need to create a new column in the data to contain the transformation.

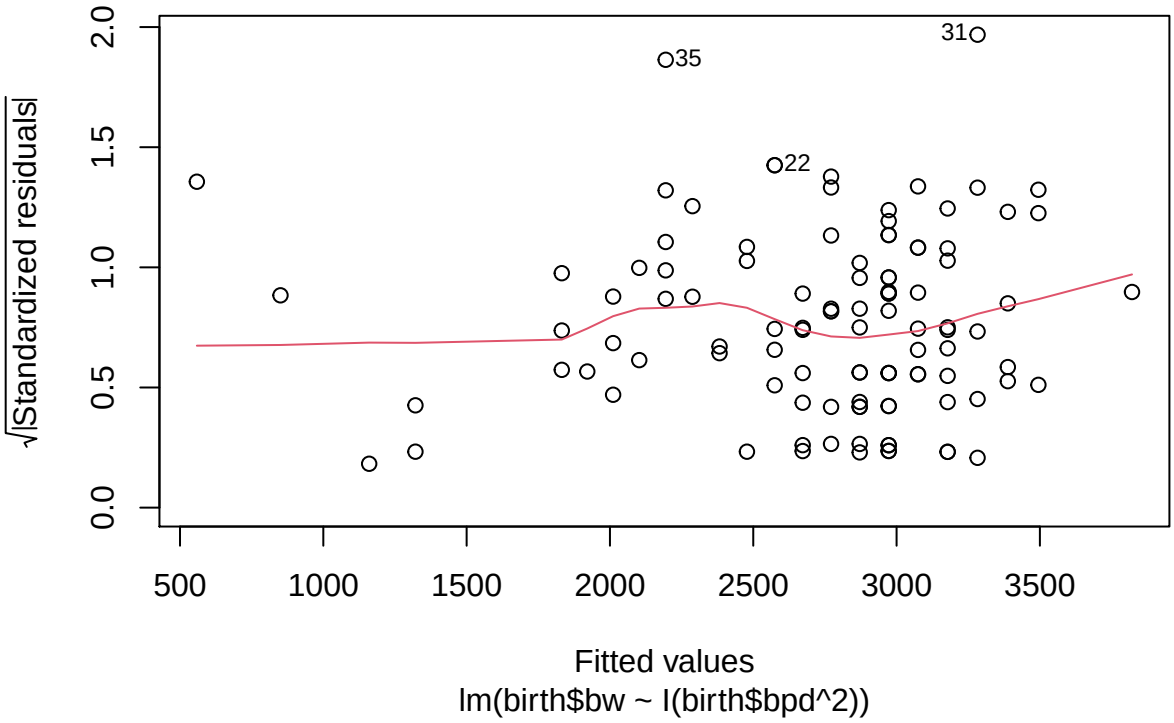
```r
bpd.tr.lm<-lm(birth$bw~I(birth$bpd^2))
summary(bpd.tr.lm)
```

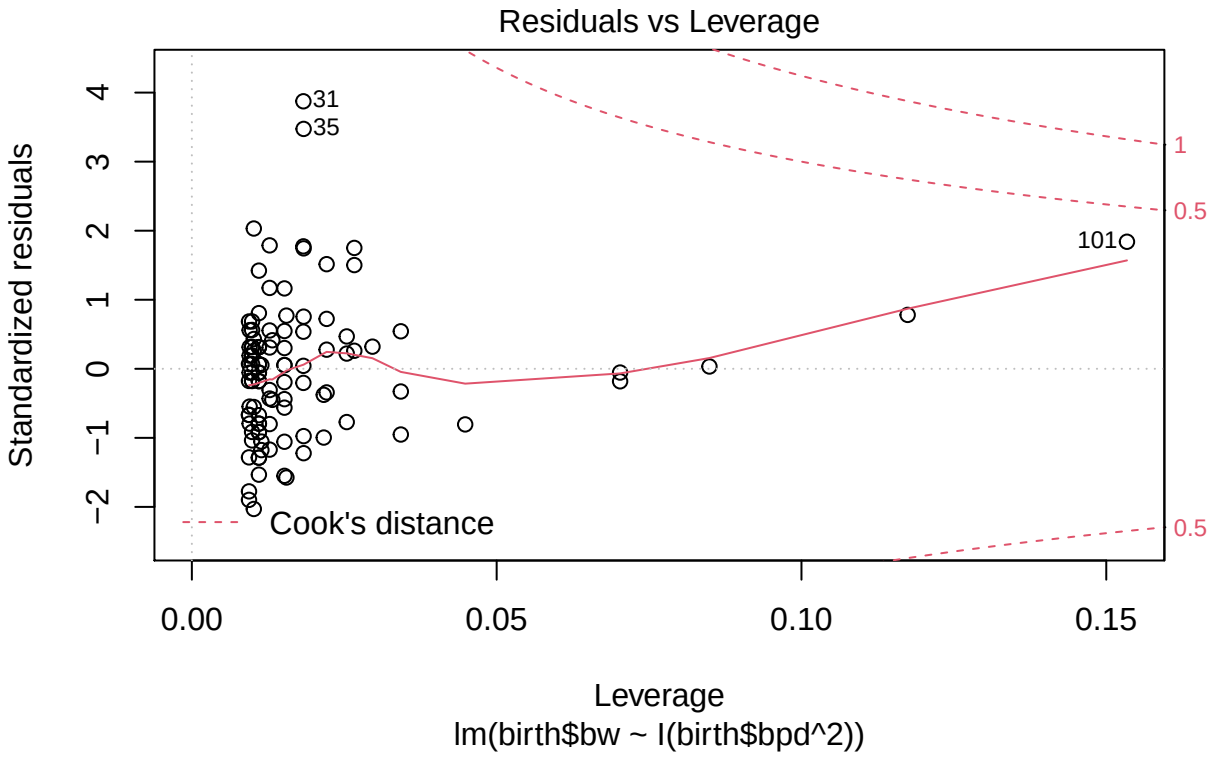
```
##
## Call:
## lm(formula = birth$bw ~ I(birth$bpd^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -824.70 -291.81   13.01  182.14 1567.39
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.705e+03  3.183e+02  -5.355 5.07e-07 ***
## I(birth$bpd^2)  5.526e-01  3.928e-02  14.068  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 408.4 on 105 degrees of freedom
## Multiple R-squared:  0.6534, Adjusted R-squared:  0.6501
## F-statistic: 197.9 on 1 and 105 DF,  p-value: < 2.2e-16
```

```r
plot(bpd.tr.lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(birth$bw ~ I(birth$bpd^2))
31
35
22
−2
−1
0
1
2
3
4


Scale−Location
√|Standardized residuals|
Fitted values
lm(birth$bw ~ I(birth$bpd^2))
31
35
22
0.0
0.5
1.0
1.5
2.0
500
1000
1500
2000
2500
3000
3500



This transformation has improved the situation marginally. The r^2 has increased and the trend in the residuals is less strong (but not completely gone). We may want to try a more complex model that uses both the bpd and transformations of it. (We will discuss this when we cover multiple regression)

(7) An expectant mother has been told their baby's bpd is 80 - what is the estimated birthweight?

Explain your answer, including which model you used and why

The second model should be used as it does map the relation between the attribute we have a value for and our dependent variable (BW).

```
bpd.lm
```

```
##
## Call:
## lm(formula = birth$bw ~ birth$bpd)
##
## Coefficients:
## (Intercept)    birth$bpd
##    -5505.41        92.14
```

From the model coefficients we can see that the relation between bw and bpd is:

$$bw = -5505.41 + 92.14 \times bpd$$

So to compute this:

```
bw.80<--5505.41+92.14*80
bw.80
```

```
## [1] 1865.79
```

An expectant mother's estimated baby birthweight is 1866 gr if her bpd is 80.

(8) An expectant mother has been told their's ad is 105 - what is the estimated birthweight?

Explain your answer, including which model you used and why

The first model should be used as it does map the relation between the attribute we have a value for and our dependent variable (BW).

```
ad.lm
```

```
##
## Call:
## lm(formula = birth$bw ~ birth$ad)
##
## Coefficients:
## (Intercept)     birth$ad
##    -2867.92        55.12
```

So to predict a BW we can use the coefficients to write out the relationship:

$$bw = -2867.92 + 55.12 \times ad$$

```
bw.105<--2867.92+ 55.12*105
bw.105
```

```
## [1] 2919.68
```

THe estimated birth weight for an AD=105 is 2919 grams.

(9) The mean birth weight in the UK is 3300 gr, given this sample of data test this hypothesis.

The hypothesis to test can be $H_0 : \mu = 3300$ and $H_1 : \mu < 3300$ This is a one sided test as the mean in the sample is smaller than the UK one. The mean in the sample is:

```
mean(birth$bw)
```

```
## [1] 2739.093
```

The test statistic to use in this case is based on Z (as we have more than 30 measurements in our sample):

```
test.st<-(mean(birth$bw)-3300)/(sd(birth$bw)/sqrt(107))
test.st
```

```
## [1] -8.405048
```

Then we can find how likely this test statistic using Z (the standardised normal distribution)

```
pnorm(test.st)
```

```
## [1] 2.138436e-17
```

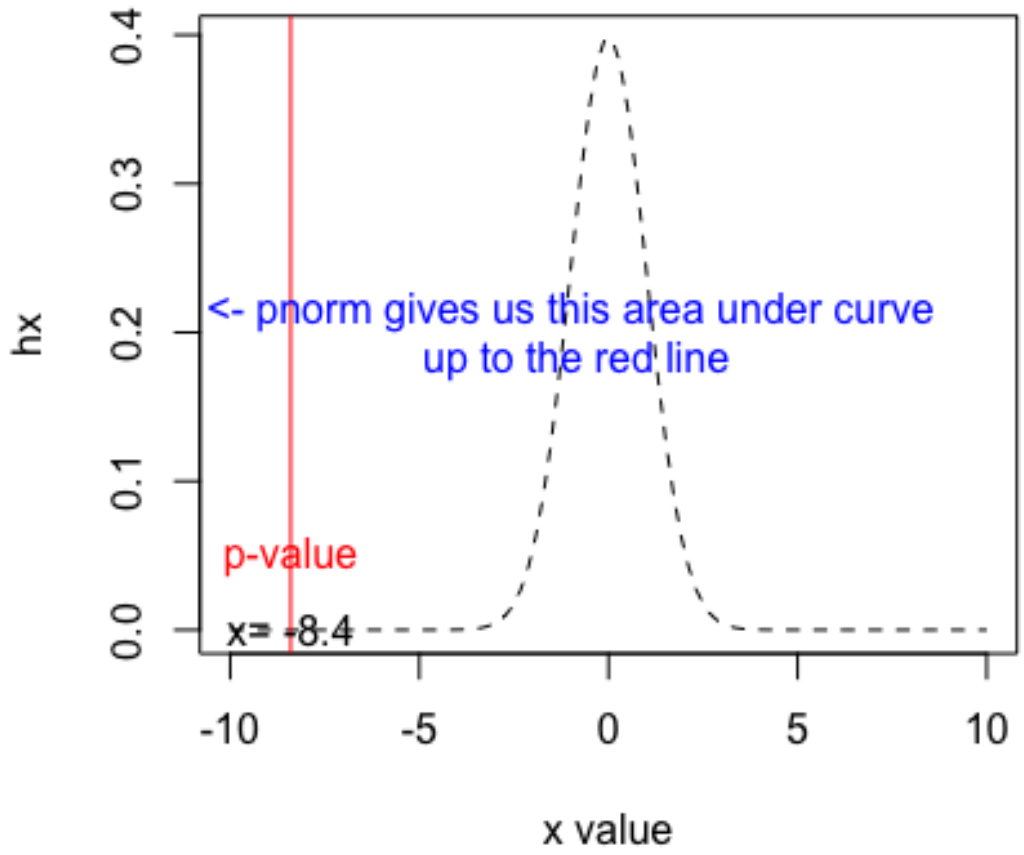


Figure 1: The test statistic on the Normal distribution

The result of pnorm(test statistic) gives us the probability of having such a value or more extreme (in this case smaller), see figure 1. Such a sample mean is very unlikely as the p-value is extremely small. Therefore based on this sample we would reject the Null Hypothesis.

In practice this wont mean that the mean birth weight for babies in the UK will be changed as there are 600 000 (approx) births a year and this sample was only containing 107 (The power of this test is not high). This result would likely lead to the need to take a much larger sample to confirm the result, and also to check whether the sample we used was in some way biased. We could have been sampling from "low weight" or clinically complex situations, that would explain this lower mean.