Outliers and Influential Observations

Irfan Kanat

August 13, 2017

Outliers and Why They Are Important

Outliers are extreme observations in a sample. The reason we care about them is that they can bias the fit quite drastically. Let me demonstrate with an example:

You remember develop dataset from Interactions learning activity. Let us say that two aliens who were visiting the planet made a mistake in their height calculations (must be the imperial system) and disguised themselves as 4mt tall (13 ft) 10 year olds.

Here I will use a smaller dataset as outliers' effects are dampened a bit with a larger dataset. One or two observations will have less of an effect in a larger dataset. **I AM REPEATING THIS AS IT IS IMPORTANT: Outliers are less of a concern in a larger dataset.**

```
# I am reducing the sample size so that the effect of outliers is more pronounced.
developSmall <- develop[1:30, ]
# Adding the aliens
developOut <- rbind(develop[1:30, ], data.frame(gender = 0:1, age = c(10:10),
                                                height = c(399, 400)))
# Fitting the model with reduced sample
developSmall_lm_0 <- lm(height ~ gender + age, data = developSmall)
# Fitting the model with aliens
developOut_lm_0 <- lm(height ~ gender + age, data = developOut)
```

Let us compare the two fits. (I will use screenreg to print two models side by side, you will need texreg package for this).

```
screenreg(list(developSmall_lm_0, developOut_lm_0))
```

```
##
## ================================
##              Model 1    Model 2
## --------------------------------
## (Intercept)   8.03      120.83
##              (6.12)     (61.35)
## gender       19.53 ***   26.09
##              (2.51)     (25.74)
## age           8.70 ***    1.29
##              (0.45)      (4.55)
## --------------------------------
## R^2           0.95        0.04
## Adj. R^2      0.94       -0.03
## Num. obs.    30          32
## RMSE          6.80       72.03
## ================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

```
summary(developSmall_lm_0)$fstatistic # Compare F statistics
```

```
##    value    numdf    dendf
## 232.6453   2.0000  27.0000
```

```
summary(developOut_lm_0)$fstatistic
```

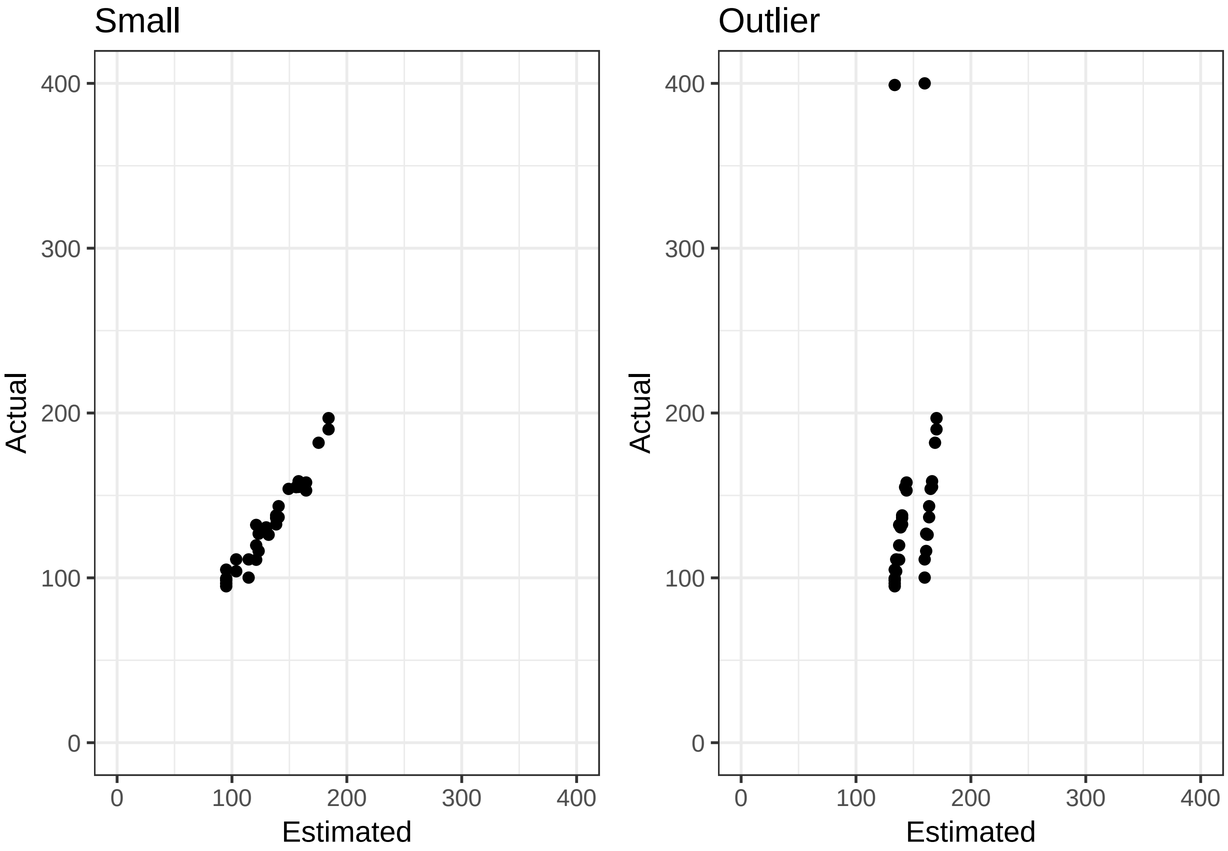
```
##      value      numdf      dendf
##  0.5775267  2.0000000 29.0000000
```

You will see that our variable estimates turned insignificant. The model's explanatory power has suffered. Furthermore, looking at F statistic, the model with outliers is basically unusable.

Let us visually inspect how the outlier model compares to small model.

First off, fitted values to actual values for both models.

```
# Easy to remember, easy to follow
plotDevSmall <- qplot(x = developSmall_lm_0$fitted.values,
                      y = developSmall$height, geom = "point") +
  ggtitle("Small") + xlab("Estimated") + ylab("Actual") + ylim(0, 400) +
  xlim(0, 400) + theme_bw()
# Directly uses the model, but have to remember .resid and .fitted
# plotDevSmall<-ggplot(developSmall_lm_0, aes(.fitted, .resid))+ geom_point()

plotDevOut <- qplot(x = developOut_lm_0$fitted.values,
                    y = developOut$height, geom = "point") +
  ggtitle("Outlier") + xlab("Estimated") + ylab("Actual") + ylim(0, 400) +
  xlim(0, 400) + theme_bw()

grid.arrange(plotDevSmall, plotDevOut, nrow = 1)
```

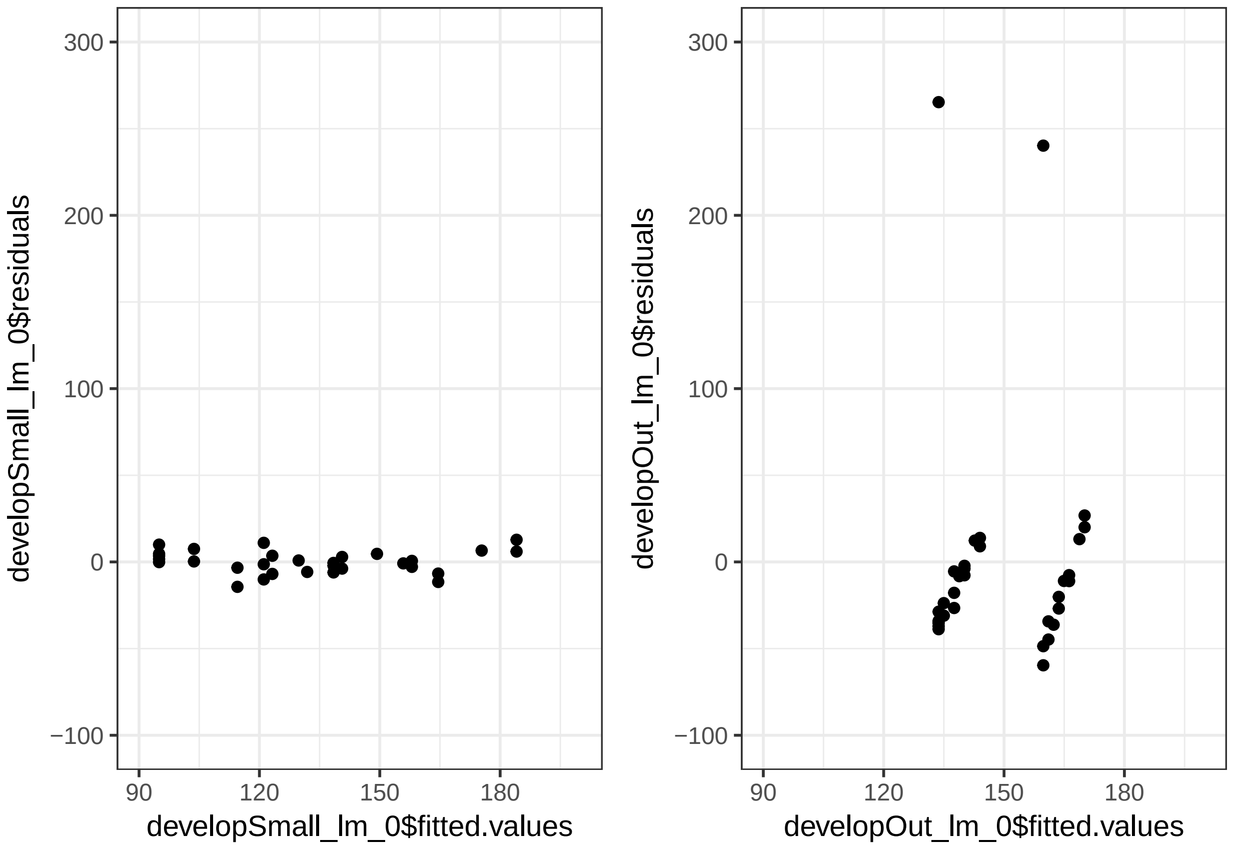


On the X axis you see our guess for what the person's height would be. Y axis is their actual height. If we made perfect predictions the points would lie on a 45 degree angle (assuming x and y axes are on same scale).

Original model is a pretty good fit. Whereas with outliers our guesses are wildly inaccurate.

How inaccurate? Let us inspect residual plots.

```
plotDevSmall <- qplot(x = developSmall_lm_0$fitted.values,
                      y = developSmall_lm_0$residuals, geom = "point") +
  ylim(-100, 300) + xlim(90, 200) + theme_bw()

plotDevOut <- qplot(x = developOut_lm_0$fitted.values,
                    y = developOut_lm_0$residuals, geom = "point") +
  ylim(-100, 300) + xlim(90, 200) + theme_bw()

grid.arrange(plotDevSmall, plotDevOut, nrow = 1)
```



You can see the model with outliers has more variance (remember less variance in residuals means more accuracy in estimates) in residuals. Also it fits the values into a narrower band.

In both plots we can clearly see the outliers. Unfortunately, in real life making the call for outliers is not as easy as spotting 13ft tall school kids.

Univariate Approaches to Outlier Detection

If the idea is extreme observations, we can use the normal distribution to guide us in detecting them. As you may have learned (when conducting T tests) 99% of the observations are expected to be 2 standard deviations around the mean. We can use this to say if an observation is beyond 3 standard deviations from the mean that would be an extreme observation.

Simply calculate the mean, and determine cut-off values based on standard deviation.

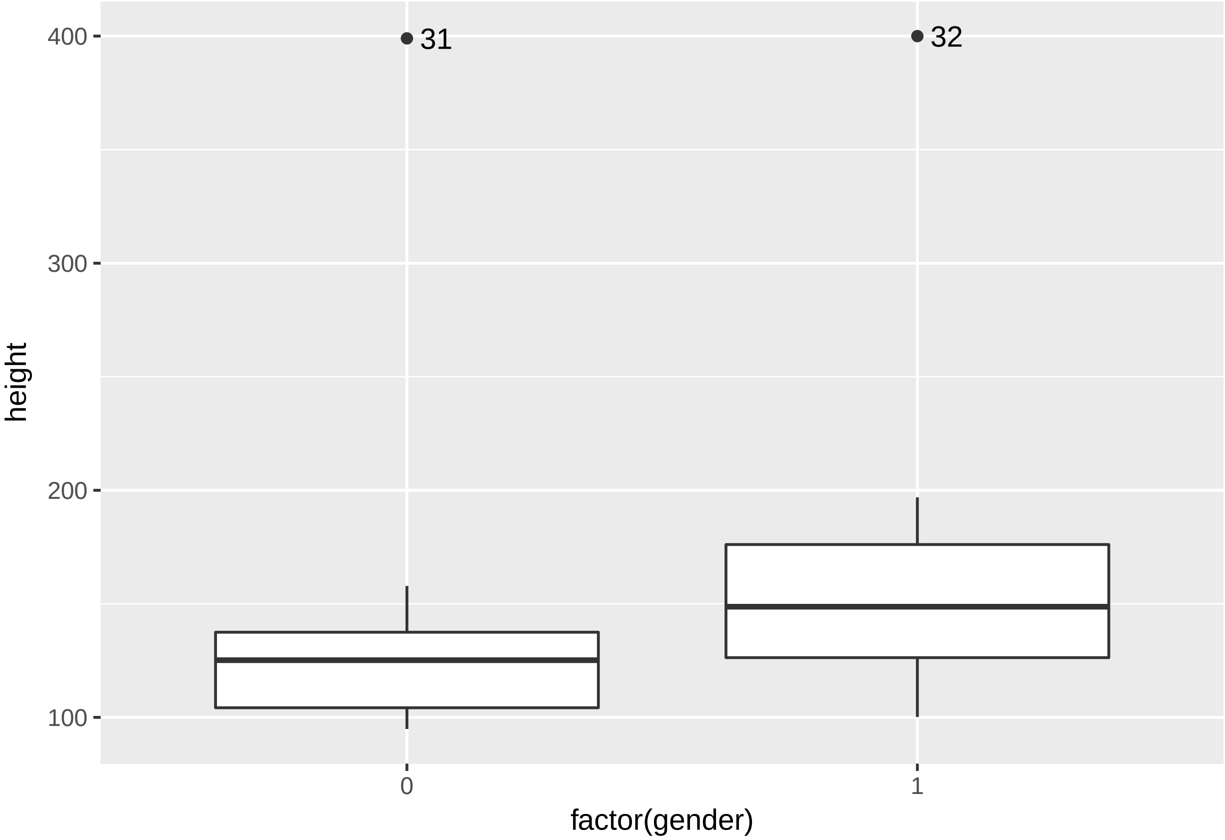
```
u <- mean(developOut$height) # Get the mean
s <- sd(developOut$height) # Get the sd

upper <- u + 3 * s # any kid taller than this would be an outlier
lower <- u - 3 * s # any kid shorter than this would be an outlier
# Filter the outliers
developOut[developOut$height > upper | developOut$height < lower, ]
```

```
##    gender age height
## 31      0  10    399
## 32      1  10    400
```

You already know how to remove observations that you filtered from Module 2. So here is a simple exercise for you: Create a new dataset from developOut that does not have the outliers in.

Visually you can inspect the boxplots.

```
# Create A Boolean (True False) Vector for Outliers to use in labeling
outlier <- (developOut$height > upper | developOut$height < lower) # True/False
outlier[outlier] <- which(outlier) # Put the row number in
outlier[outlier == 0] <- "" # Leave blank if not outlier
# Create a boxplot
qplot(data = developOut, x = factor(gender), y = height, geom = "boxplot") +
  # Label the outliers by row numbers
  geom_text(aes(label = outlier), na.rm = T, hjust = -0.4)
```



You can change the criteria. I used 3 standard deviations. That assumes a normal distribution. Instead you can use quartiles for example.

Here is a simple exercise for you, I used row numbers to identify outliers. Can you print the height of the outliers instead?

Multivariate Approaches to Outlier Detection

It is good to know if certain observations are extreme or not, but do they really have an extraordinary impact in your model? You may discard datapoints because they are different, but what if the variable they were abnormal on was not that significant in your model anyway?

Wouldn't it be nice to have a test that is specific to the model you are fitting? Say if an observation is changing the model fit more than others? If an observation is more influential?

Cook's Distance

With Cook's Distance, we compare mean estimate when an observation is present versus when an observation is removed. How much does the overall estimation change when we remove a variable? For an influential observation, we expect this change (distance) to be large.

Here is how we calculate cooks distance in R:

```r
cooks.distance(developOut_lm_0) # This gives us a distance calculated for each observation.
```

```
##            1            2            3            4            5
## 5.710332e-03 1.024964e-02 8.115103e-03 2.972097e-03 7.658995e-04
##            6            7            8            9           10
## 3.409174e-02 2.950912e-04 7.934888e-05 5.774748e-03 9.623946e-03
##           11           12           13           14           15
## 1.468093e-02 8.684442e-03 2.744809e-03 5.050972e-03 1.160421e-03
##           16           17           18           19           20
## 2.694476e-05 1.187956e-04 2.264304e-02 2.180004e-03 7.825378e-03
##           21           22           23           24           25
## 1.806029e-03 1.271558e-03 3.632287e-04 3.875848e-03 1.054918e-02
##           26           27           28           29           30
## 8.610571e-03 6.562473e-04 2.827888e-03 1.531919e-03 3.212384e-04
##           31           32
## 4.927992e-01 5.531347e-01
```

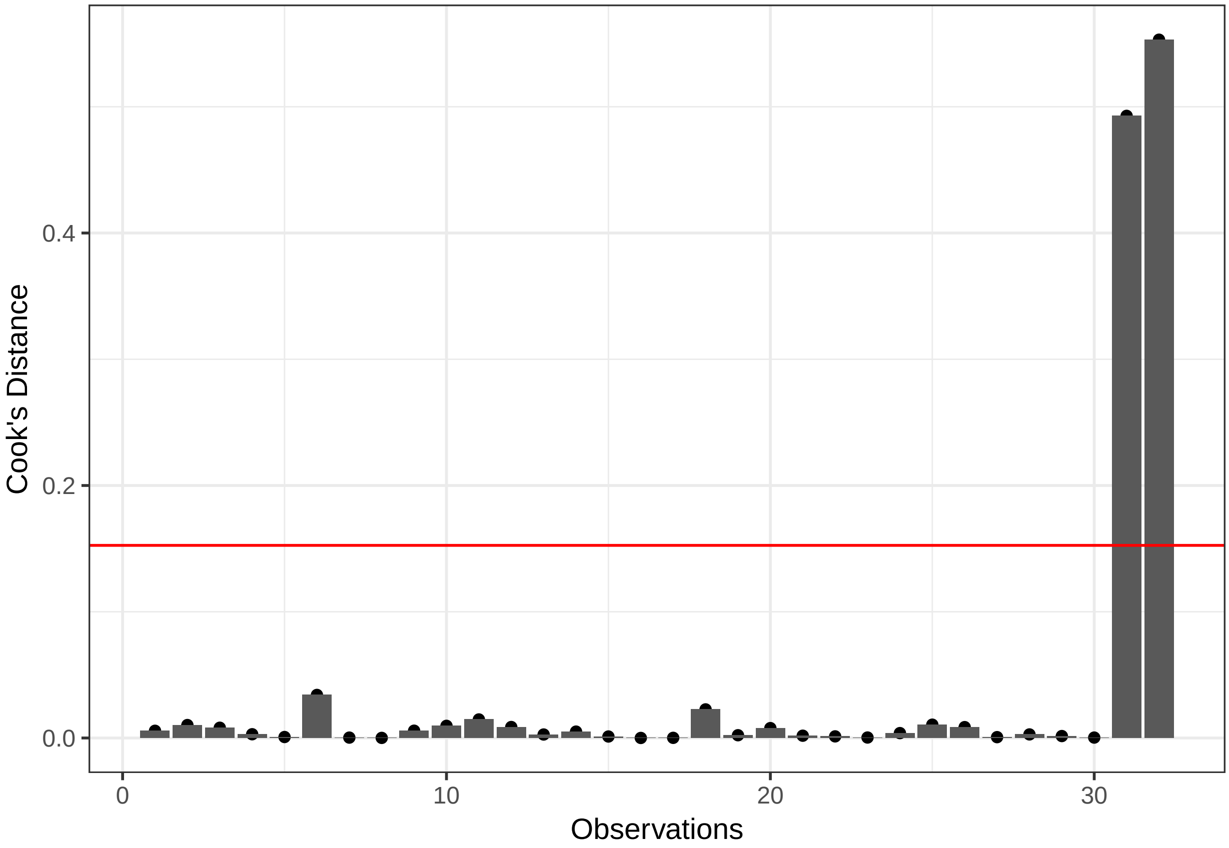
Generally speaking we would assume an observation is an outlier if the cooks distance is greater than 4 times the average cooks distance for the model.

```r
cd <- cooks.distance(developOut_lm_0) # Save distances in an array
cd[cd > mean(cd) * 4] # Filter the distances that are greater than 4 times the mean distance.
```

```
##        31        32
## 0.4927992 0.5531347
```

Let us visualize

```r
qplot(y = cd, x = seq_along(cd)) + geom_bar(stat="identity") +  # Easy to remember
  geom_hline(yintercept = mean(cd * 4), col="red") + # Add a line at 4 times the mean of cook's distanc
  xlab("Observations") + ylab("Cook's Distance") + theme_bw()
```



```
#ggplot(developOut_lm_0, aes(seq_along(.cooksd), .cooksd))+ # Directly uses fitted model
# geom_bar(stat="identity", position="identity")+
# geom_hline(yintercept=mean(cd*4),col="red")+
# xlab("Observations") + ylab("Cook's Distance") + theme_bw()
```

Exercises

1 - Create a new dataset from developOut that does not have the outliers in it.

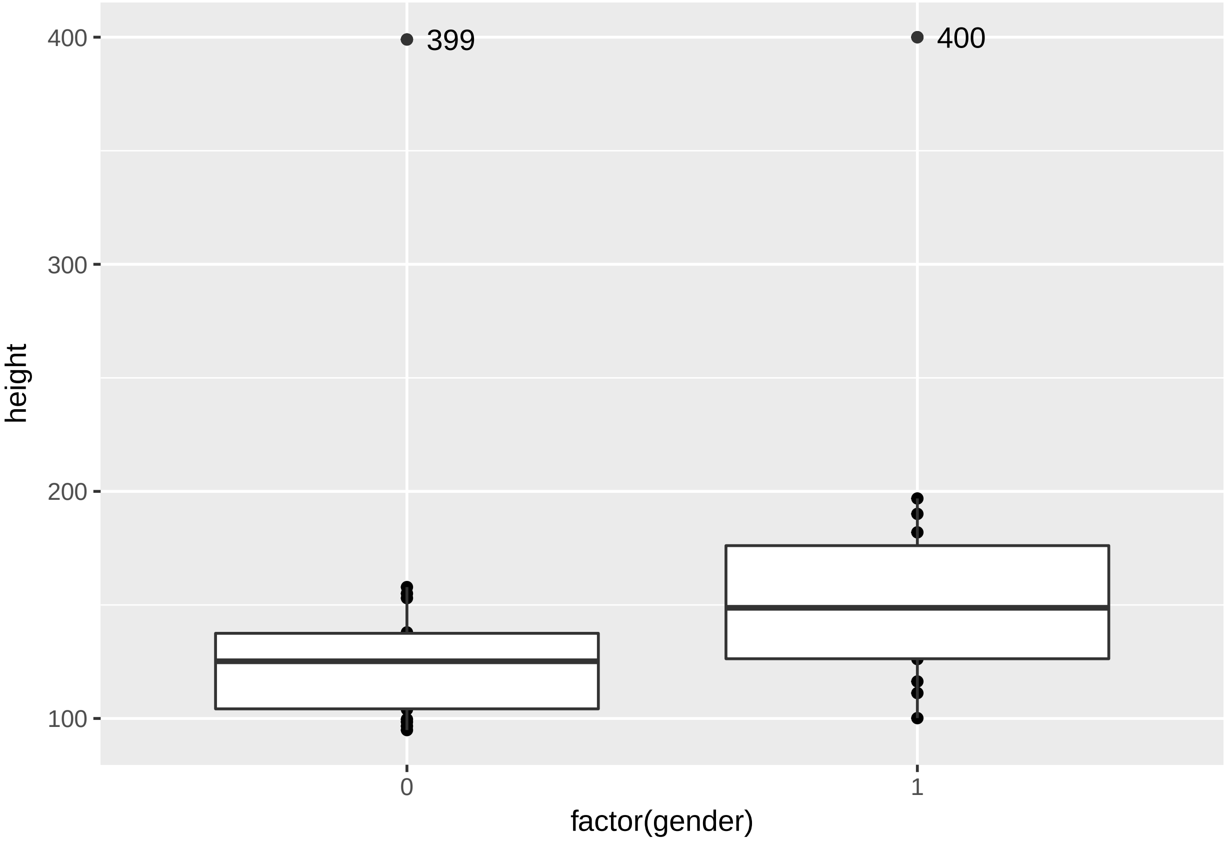
```
developIn <- developOut[!(developOut$height > upper | developOut$height < lower), ]
tail(developIn)
```

```
##    gender age    height
## 25      0  10  94.88515
## 26      1  11 126.78253
## 27      1  14 153.97502
## 28      0  13 110.97950
## 29      0  17 155.02829
## 30      0  15 132.42856
```

2 - I used row numbers to identify outliers. Can you print the height of the outliers instead?

```
# Create A Boolean (True False) Vector for Outliers to use in labeling
outlier <- (developOut$height > upper | developOut$height < lower) # Outlier?
outlier[outlier] <- developOut[outlier, "height"] # Put the row number in
outlier[outlier == 0] <- "" # Leave blank if not outlier
# Create a boxplot
```

```
qplot(data = developOut, x = factor(gender), y = height) + geom_boxplot() +
  # Label the outliers by height
  geom_text(aes(label = outlier), na.rm = T, hjust = -0.4)
```


height
factor(gender)