STAT0006 ICA 3

Group 19

Student numbers: 20049301, 19014604, 19184272, 20022528. Word Count : 1995

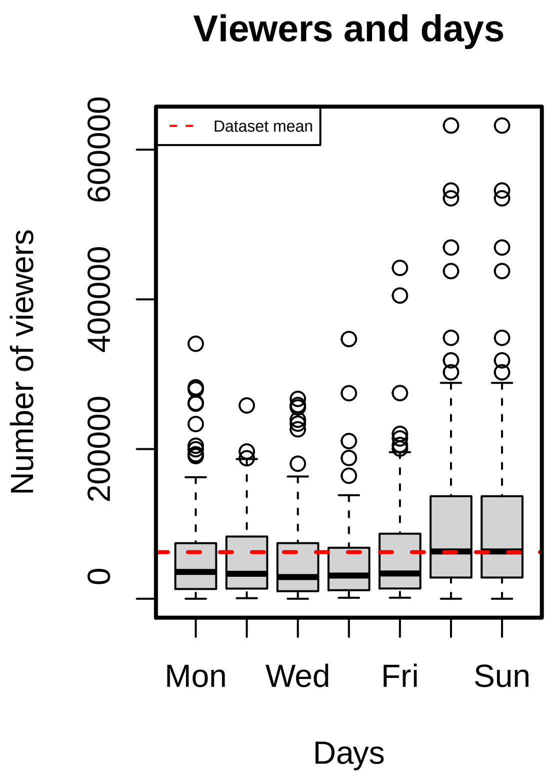
Contents

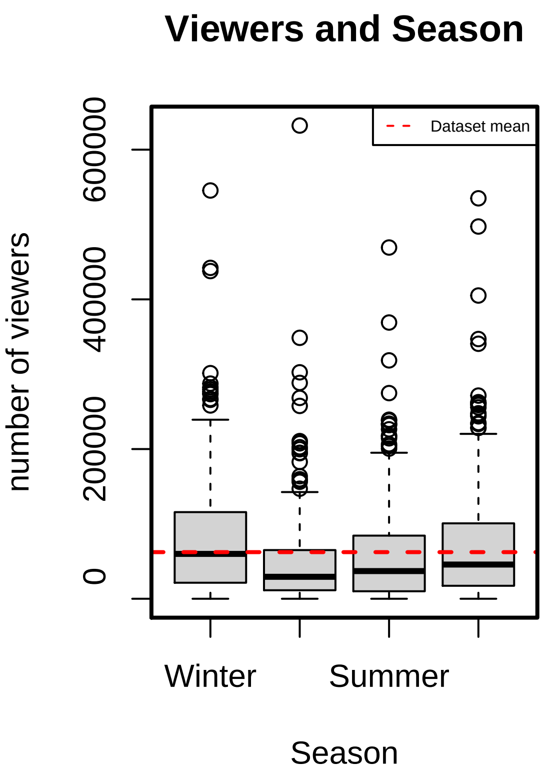
Introduction to the data

A group of five medias personalities running an online streaming channel want to understand in which ways various variables influence their number of viewers. To do so, they recorded the values of those variables during their three years of live streams, providing at the end a dataset containing 9 variables, which gives, in total, results for 1542 live streams. There are no missing data.

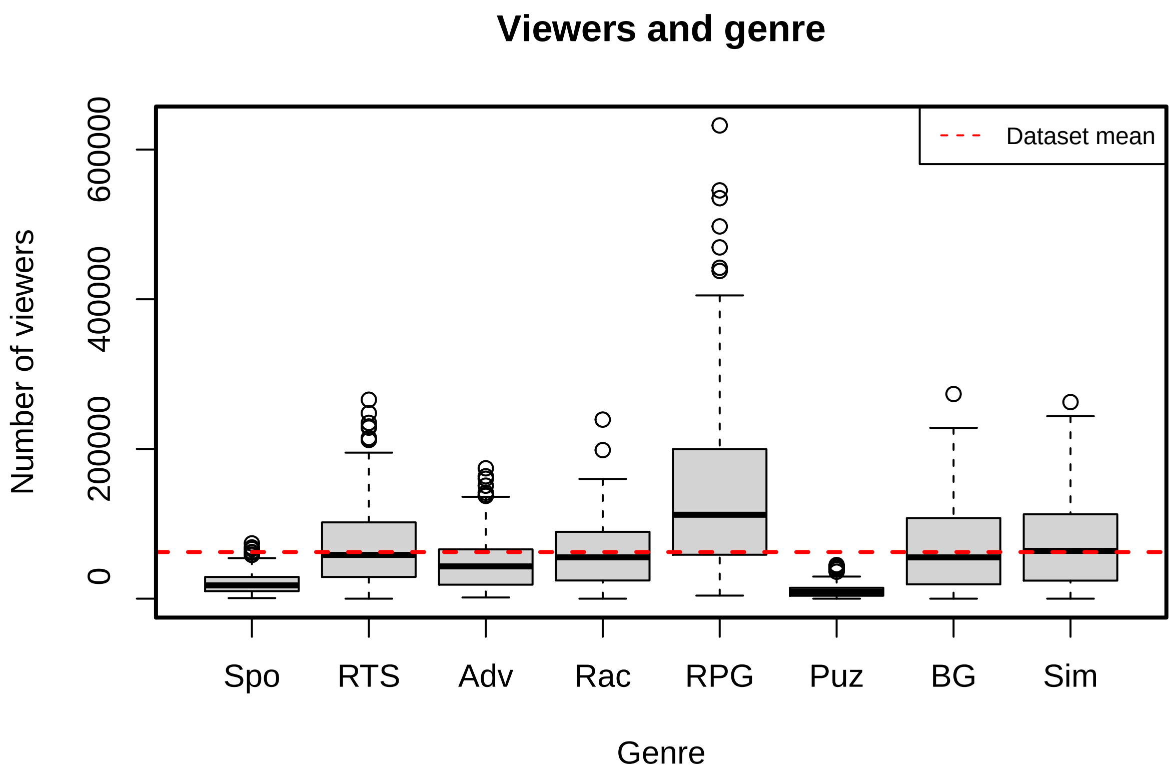
For each of the 1,542 live streams, the players have recorded : the total amount of viewers who watched the stream, the genre of video game played (i.e. Puzzle, Racing, . . .), the host (which one of the 5 influencers played), the amount of suscribers of the chanel at the time of the live, the day of the week, the season of the year, the number of guest players included in the live stream, and finally the number of adverts both in the actual and previous live stream.

We investigate the influence of some of those variables using the comparative boxplots below since the number of observation for the levels of the categorical covariate is approximately the same. The datasest mean of the amount of viewers, whose value is 62,264, has been provided for each boxplot.



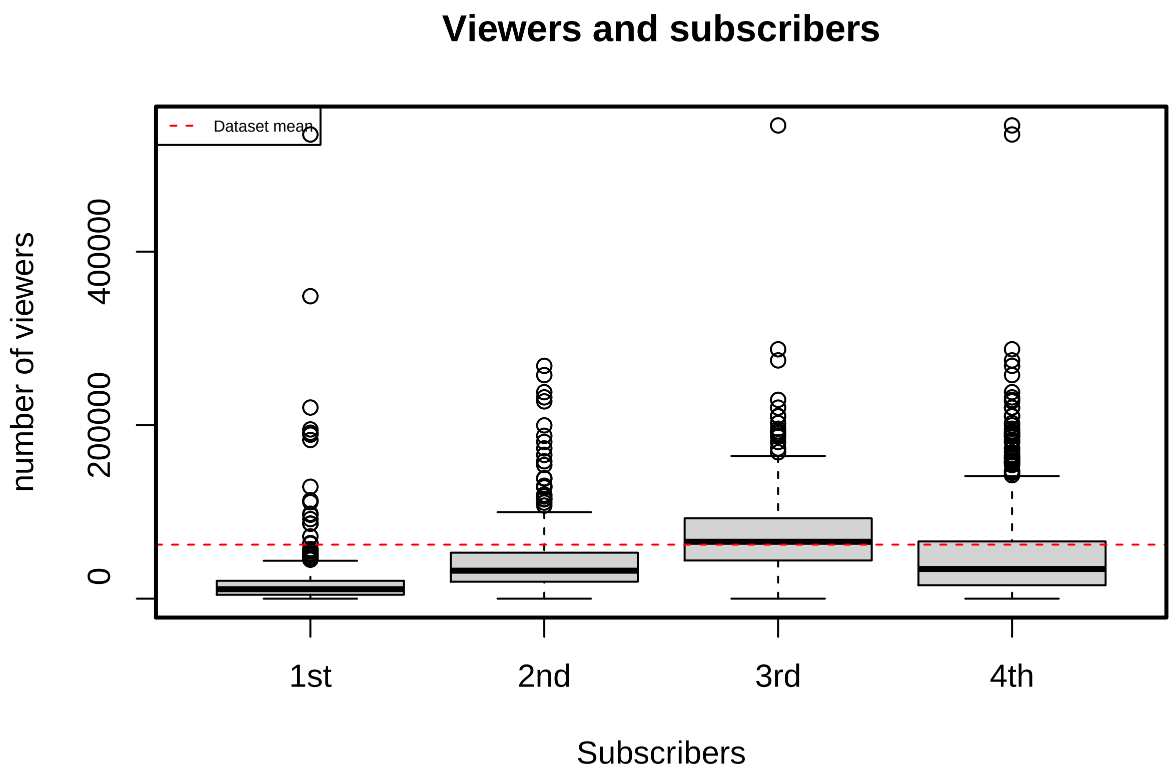


First, for ‘days’, we remark a real disparity. The number of viewers tends to largely increase during weekends, where the number of viewers can be up to two times larger than during the week. Similar differences are observed for the variable ‘season’, for instance between winters and springs.



Regarding the 'genre of video game', we note more flexibility as role playing games, for example, tend to be far more popular than puzzle or sports games. The maximum number of viewers puzzle games attained is approximately at the same level that the first quartile of role playing games[1].

[1] 'Spo' = Sports; 'RTS' = Real Time strategy; 'Adv' = Adventure; 'Rac' = Racing; 'RPG' = Role Playing games'; 'Puz' = Puzzle; 'BG' = Board Games; 'Sim' = Simulation



The data being roughly ordered in ascending order of the number of subscribers[2], this variable is also positively correlated with the number of viewers.

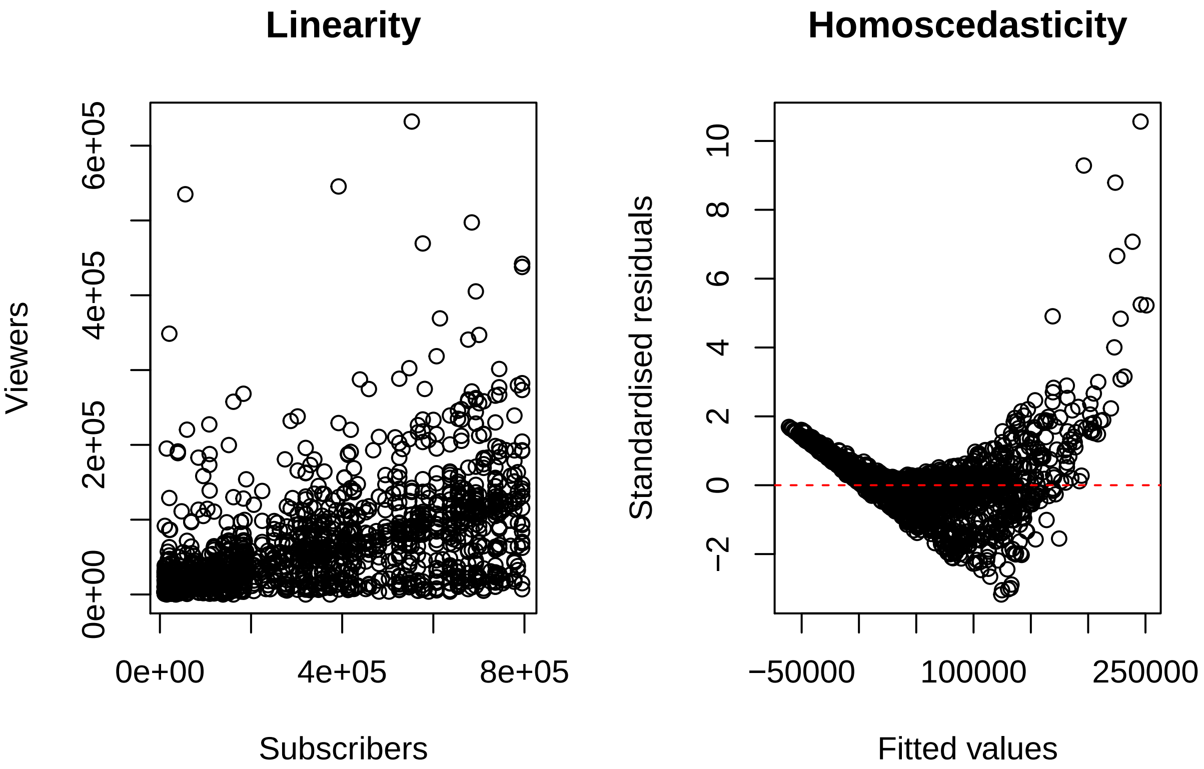
In conclusion, various covariates such as ‘season’, ‘genre’ or ‘subscribers’ may be considered when building a model for the number of viewers.

[2] Data separated in four parts. ‘1st’ = 1-385, ‘2nd’ = 386 - 771, ‘3rd’ = 771 - 1157, ‘4th’ = 1157 - 1542

Model building

Our first step in this process is to construct a normal linear model, without any modification in order to check what we have to improve. The second step consists in using various tools such as interactions or transformations to improve the model progressively.

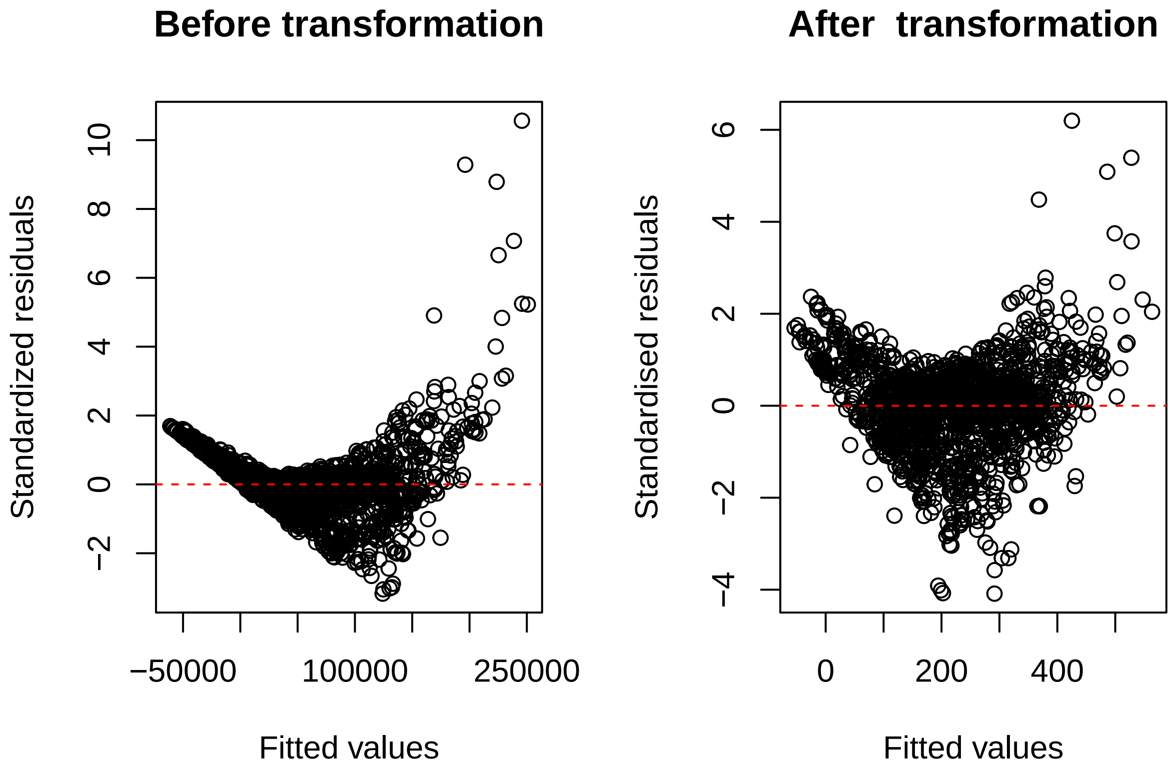
Our issues



To precise our method, we checked linearity by plotting the relationship 'subscribers/viewers' specifically in the model as the three other numeric covariates only have four values, with a number of observation different for each value. And as observed, this relationship seems to be linear but can be improved.

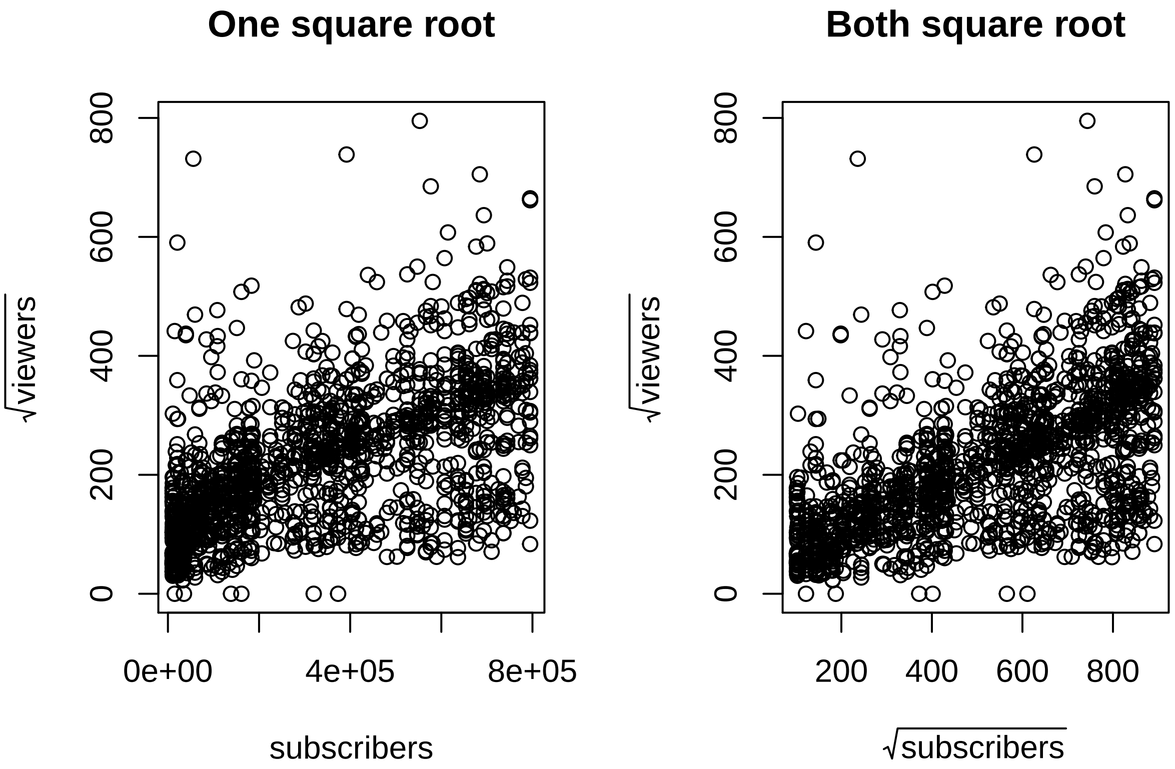
Regarding the homoscedasticity assumption, the model fits negative values and the points form a real quadratic shape. Therefore, it is largely detrimental. However, it suggests intuitively a square root transformation on the response variable.

Transformations



No negative fitted values are observed anymore, and our coefficient of determination increased by 0.10, which is a good improvement. But progress can still be made as we still observe a quadratic shape.

However, this transformation eases the interpretation of variables. So we will continue to use it for now, while trying to improve the model with other methods. A change in the transformation for a more complex method, in this case with a smaller root, will be done if needed.



Since we did a square root transformation on the response variable, this logically implies the same transformation on subscribers as their values range approximately on the same scale. As we can see, the two plots seem to be the same, but one really big scatter of points remains at the bottom-left of the first plot, meaning that we will have more problems to predict large values. Hence, we decide to implement a second square root transformation for subscribers.

We do not consider a transformation on the other numeric covariates as, said earlier, their values range only between 0 and 3, which changes the y-absis values without changing the look of the plots. As we do not have other numerical covariates, the next step in our modeling process is to include interactions.

Interactions

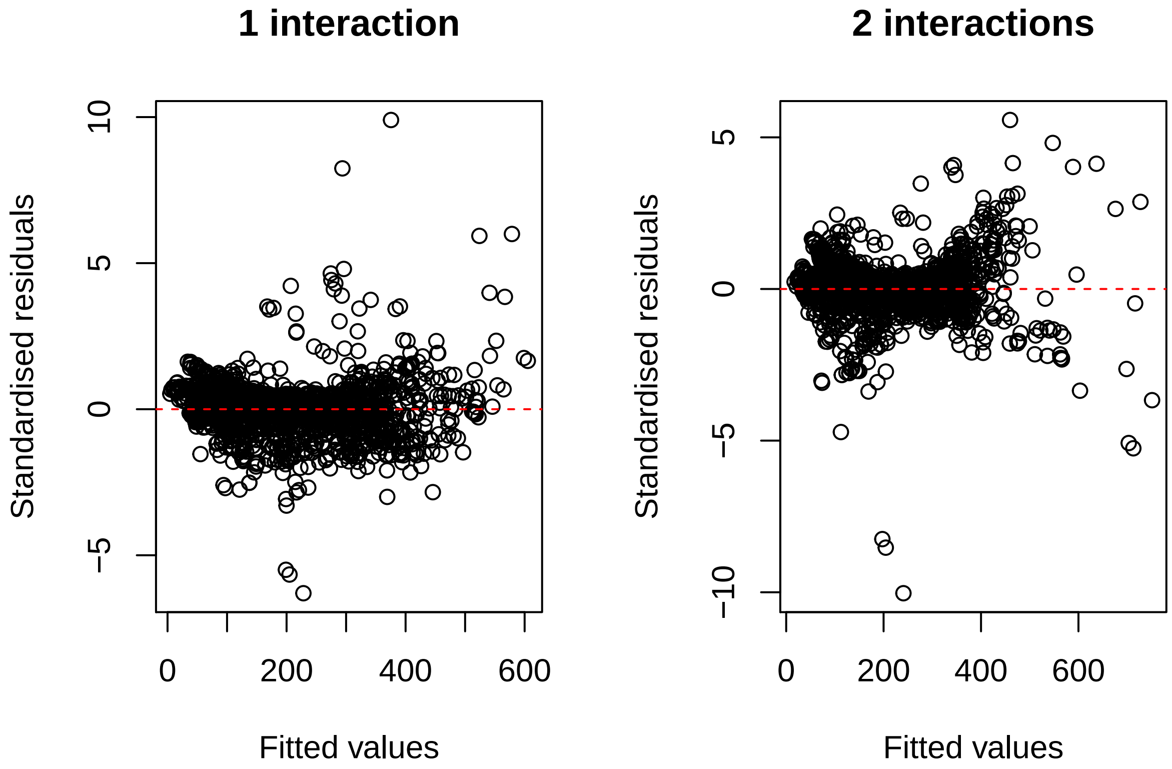
For the next part, we tried to implement the two interactions that seemed logical to us, namely the ones between 'genre' and 'subscribers' or 'guests'. In fact, increase in the number of subscribers is likely to change with the genre. Moreover, some genre are meant for multiplayer games, whilst others are meant for solo play.

What we observed is that the 'subscribers' interaction was improving the fit of the assumptions plot, which is not the case of the interaction involving the 'guest' variable. Yet, both interactions were improving significantly the coefficient of determination. So, we are sure to include the first one, nevertheless, there is a possibility that including both of them could add something to our model.

We are then left with two options: either to include only the first interaction between 'subscribers' and 'genre' (model 1), or include the two interactions (model 2) which adds 7 covariates and complexify the final interpretation.

The fundamental purpose being, at the end, to find the right balance between simplicity and precision, our next part will help us to choose the most appropriate model between the two.

Chosing between model 1 and model 2



Just by looking at the homoscedasticity assumption, we can see that the standardised residuals of the two interactions model form a real quadratic shape, which is not really the case for the second.

Leverages and outliers

Also, when analysing the leverages and outliers of the models, both of the plots contain their unusual values that could be potential outliers. But, taking the diagonal values of the hat matrix above 2*p/n[3], the two interactions model contains 77 possible leverages, while the 'one interaction' does not contain any.

Investigating if each of those 77 values is a leverage or not and then fitting the model based on this analysis can be a tedious process and can lead to withdraw lots of data.

As such, regarding the fit of the assumptions and the leverages, we decide to choose the model without the guest interaction: model 1.

Simplifying even more our model

Finally, as observed in the basic model, the three levels of the covariate season have p-values above 0.55, which suggest that we could withdraw them. And as expected, neither the coefficient of determination nor the fit of the assumptions have been impacted when doing so.

Additionally, since 'ads_last' is constructed based on 'ads_now', this suggests a collinearity between those covariates. This is further verified when conducting the variance inflation factor of the model as both variables have value equal approximately to 3.2.

Thus, we decide to withdraw the covariate with the greatest p-value, namely 'ads_now', as well as the categorical covariate 'season'.

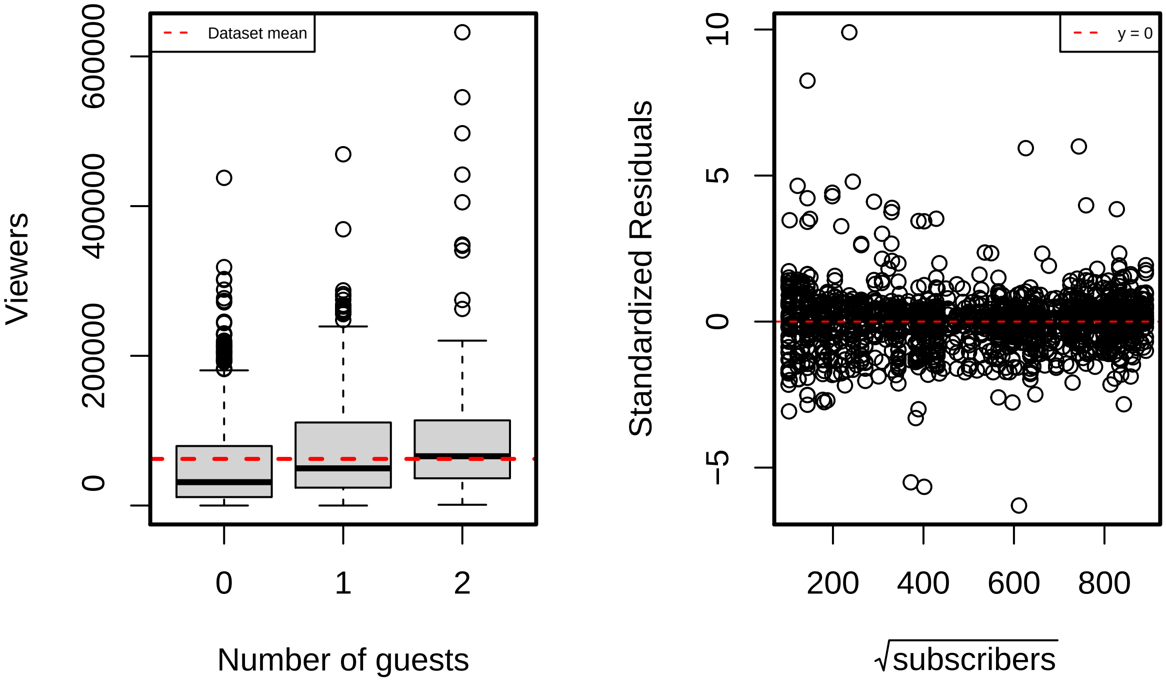
[3] 'p' : number of covariates; 'n' : number of observations

Model checking for final chosen model

```
##
## Call:
## lm(formula = sqrt(viewers) ~ genre + host + sqrt(subscribers) +
##     day + guests + ads_last + sqrt(subscribers) * genre, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -228.17  -15.34    1.06   15.73  356.04
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           35.07222    6.71156   5.226 1.98e-07 ***
## genreBoardGame                       -42.48459    8.45584  -5.024 5.65e-07 ***
## genrePuzzle                          -23.43571    8.87497  -2.641 0.008360 **
## genreRacing                          -14.96308    8.56553  -1.747 0.080858 .
## genreRealTimeStrategy                 -4.56341    8.45919  -0.539 0.589648
## genreRolePlaying                      38.85710    8.53055   4.555 5.66e-06 ***
## genreSimulation                      -31.13061    8.72774  -3.567 0.000373 ***
## genreSports                            8.56417    8.27786   1.035 0.301027
## hostPlayer2                            0.70925    2.92291   0.243 0.808308
## hostPlayer3                           30.19226    2.87524  10.501  < 2e-16 ***
## hostPlayer4                           26.73485    2.95514   9.047  < 2e-16 ***
## hostPlayer5                            3.44089    2.93363   1.173 0.241016
## sqrt(subscribers)                      0.28293    0.01188  23.815  < 2e-16 ***
## dayMon                               -19.45985    3.42069  -5.689 1.53e-08 ***
## daySat                                70.60701    3.40735  20.722  < 2e-16 ***
## daySun                                22.82055    3.43941   6.635 4.50e-11 ***
## dayThu                               -16.81111    3.55474  -4.729 2.46e-06 ***
## dayTue                               -16.82600    3.47990  -4.835 1.47e-06 ***
## dayWed                               -20.64931    3.42344  -6.032 2.04e-09 ***
## guests                                38.86019    1.45205  26.762  < 2e-16 ***
## ads_last                              -3.29219    0.86605  -3.801 0.000150 ***
## genreBoardGame:sqrt(subscribers)       0.14313    0.01483   9.653  < 2e-16 ***
## genrePuzzle:sqrt(subscribers)         -0.16998    0.01549 -10.975  < 2e-16 ***
## genreRacing:sqrt(subscribers)          0.07709    0.01498   5.146 3.01e-07 ***
## genreRealTimeStrategy:sqrt(subscribers)  0.10076    0.01488   6.772 1.82e-11 ***
## genreRolePlaying:sqrt(subscribers)     0.21428    0.01515  14.142  < 2e-16 ***
## genreSimulation:sqrt(subscribers)      0.13738    0.01504   9.134  < 2e-16 ***
## genreSports:sqrt(subscribers)         -0.15524    0.01451 -10.699  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.46 on 1514 degrees of freedom
## Multiple R-squared:  0.9123, Adjusted R-squared:  0.9108
## F-statistic: 583.5 on 27 and 1514 DF,  p-value: < 2.2e-16
```

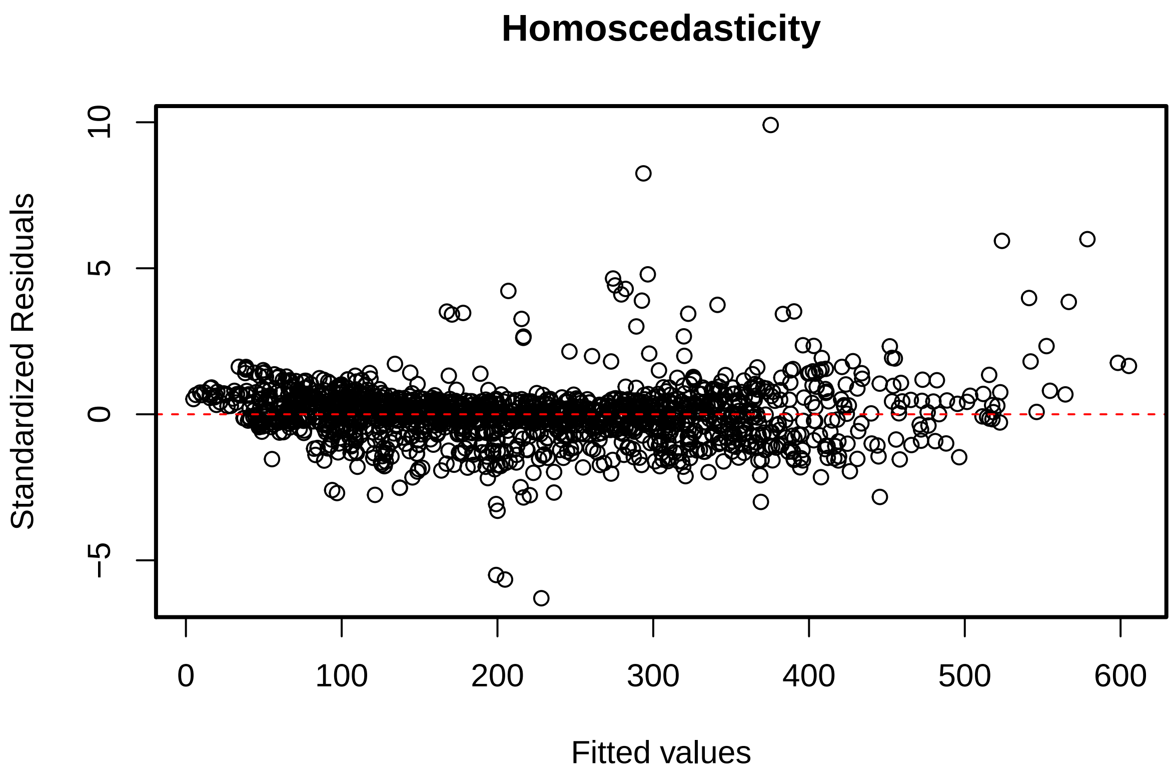
Our final model has a coefficient of determination of 0.9123, which suggests that about 91% of the variability in the number of viewers is captured by our 28 covariates.

Linearity



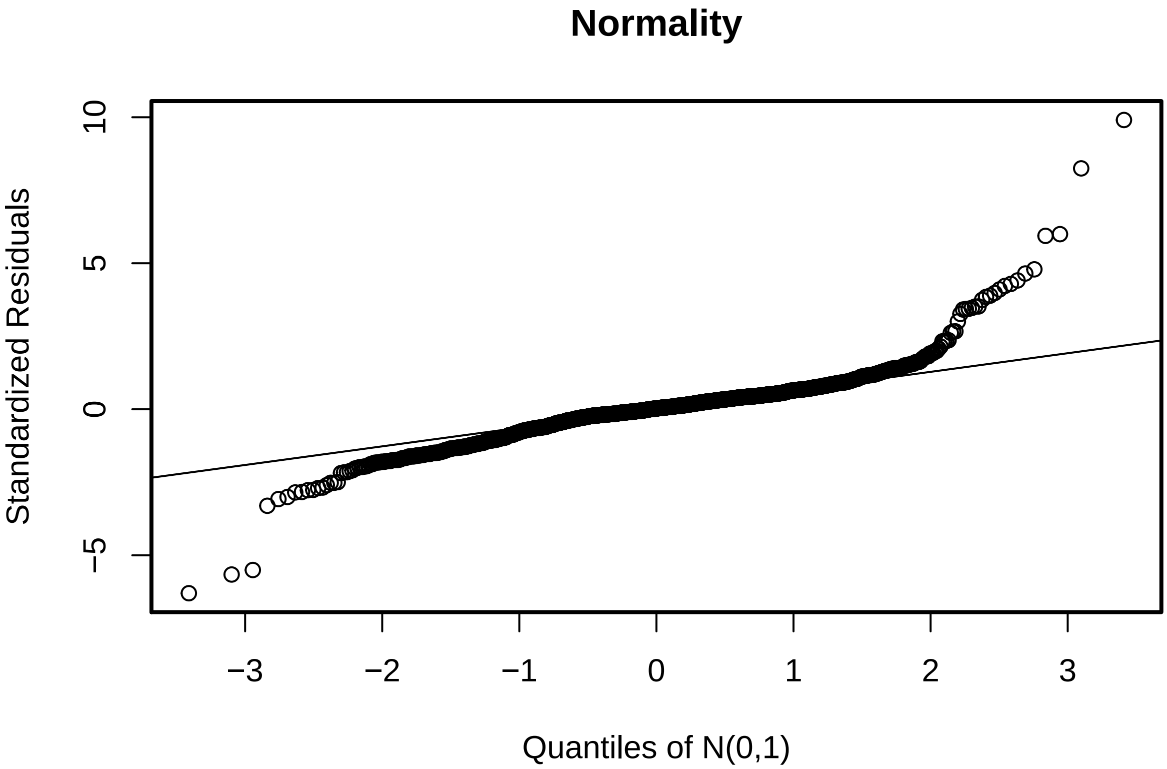
It seems that both guests and subscribers variable are positively correlated to the number of viewers. Indeed, there are some unusual standardized residuals on the second plot, but we are expecting 5% of the residuals to have a magnitude greater than 2 anyway. For the numerical variable ads_last, the relationship will be analyzed in the next section of the report as the graph contradicts the regression coefficient.

Homoscedasticity



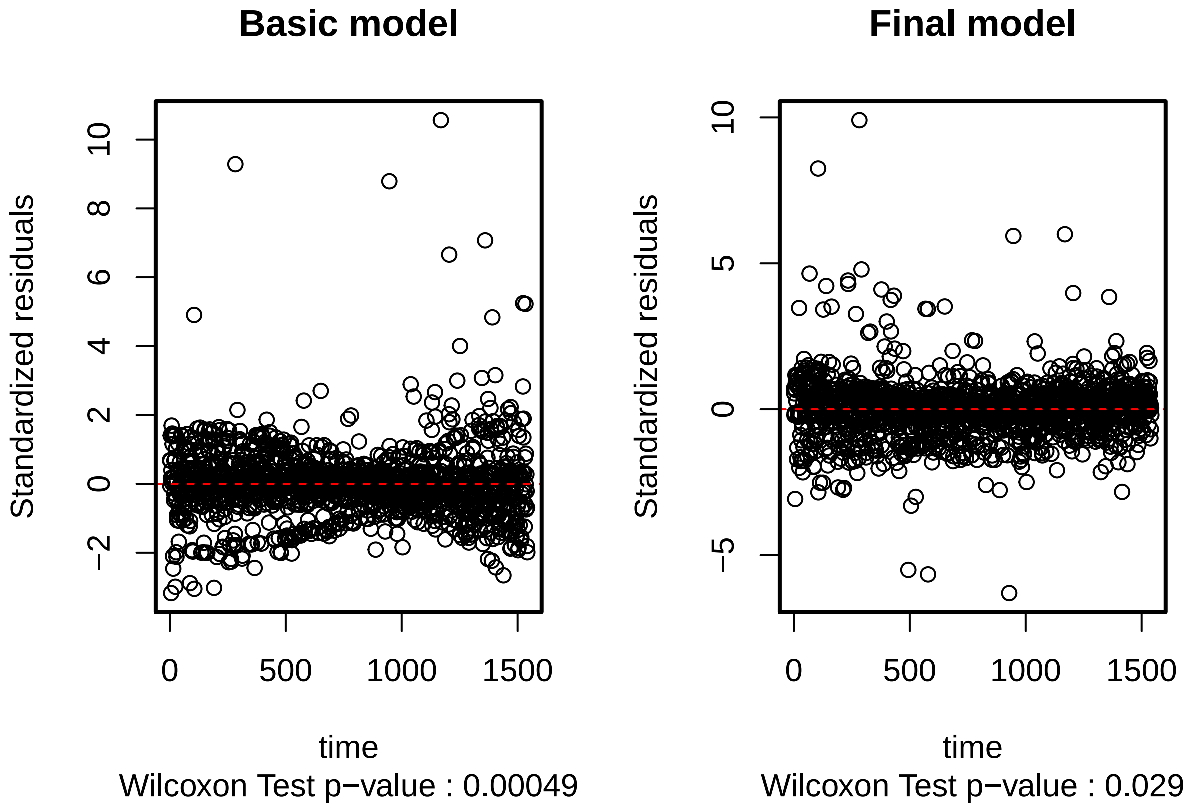
Concerning the constant variance assumption of the error terms, we can say that the plot does not look really bad. The points seem to form a random scatter points around 0 even if we observe again some unusual standardised residuals as the fitted values increase. This is logical as we are given less information for those extreme values. Something we will remark again for the next assumption.

Normality



Concerning the normality of error terms, the points from -2 to 2 seem to lie on the diagonal line. Even if values ranging from -2 to -1 lie a little bit below it. This is something that we tried to alleviate by transforming the response variable and by including interaction to improve the homoscedasticity assumption plot where many residuals were lying above 1 and in our case below -1. Regarding the tail, as we said before, we are always given less information for extreme values, which is natural.

Independence



As we know, the data are already ordered in time. Still, even if there is an improvement in comparison to the basic model as we observe on the plots, the 'Durbin Watson Test' suggests positive serial correlation as we have a 'p-value' inferior to 0.05.

Nonetheless, we cannot do many other things on our model to tackle this problem without complexifying it. Indeed, overall, the number of viewers on a live stream influences the number of viewers on the next as more subscribers are following the channel, leading to additional viewers and so on. A dependence will always remain between the data.

Conclusions

What our model tells us

Firstly, regarding the regression coefficients of the different days, four of them are negative (reference category being Friday), whereas 2 of them are positive (Saturday, Sunday), which indicates that we have more viewers during week-ends.

Following that, we also have disparities between genres of games. For instance, the expected difference in the square number of viewers between Role playing and Adventure gameplays, when the number of subscribers is set to 0 and while holding all other covariates fixed, is expected to be 38.8. Whereas this difference between Board Game and Adventure games in the same condition is expected to be -42.48. Additionally, the interaction regression coefficients all have p-values under 10^{-7}, which implies that the dependence of the number of viewers to subscribers does change with the genre of video game played, indicating that genre plays an even more important role in predicting the outcome.

For the variable host, player 2 and 5 have a week regression coefficient inferior to 4 and have p-values superior to 0.20. Whilst Player 3 and 4 have p-values inferior to 2e-16, for regression coefficients both superior to 26. This suggests that host 3 and 4 may attract more viewers than the three others.

As for the three numeric covariates, in average for the genre Adventure and holding all other covariates fixed, 10 additional subscribers increase by 784 the average number of viewers. This number goes to 1444 for the variable guest and gets back to -9 for the number of ads in the previous last stream. This means that paradoxically, even if the number of ads increase with the number of subscribers and viewers, more ads in a previous live stream generally leads to fewer viewers in the next one.

Particular issues

We note that our model contains some irregularities and tend to predict badly the number of viewers when it becomes very large. Obviously, it also does not take into account qualitative aspects such as the renown of a *guest* or[4] a host, something that can have a significant impact on the number of viewers.

[4] The term ‘guest’ is a link

New channel genre

Extracting only Adventure and Racing, we are left with 189 rows.

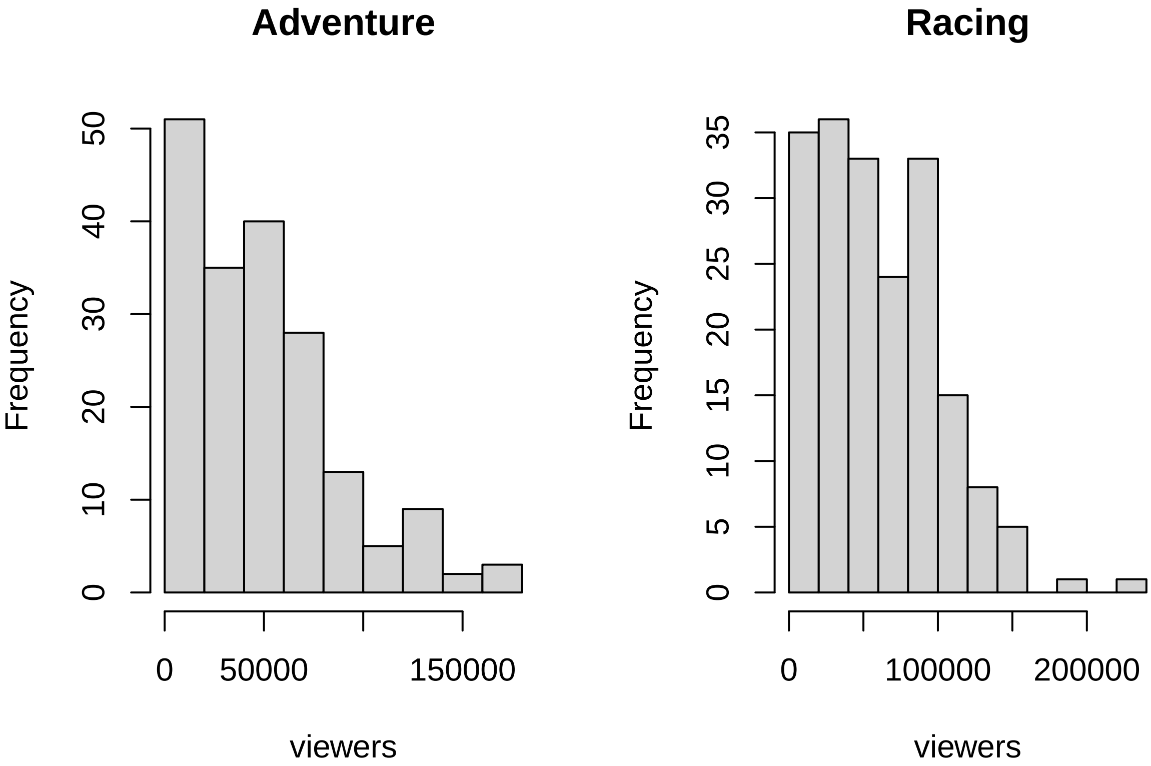
We now construct the hypothesis tests by conducting a two sample t-test :

```
##
##  Welch Two Sample t-test
##
## data:  new[new$genre == "Racing", ]$viewers and new[new$genre == "Adventure", ]$viewers
## t = 1.4819, df = 186.09, p-value = 0.14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2802.455 19724.651
## sample estimates:
## mean of x mean of y
##  57671.34  49210.24
```

We observe that the p_value calculated from the t-test is of 0.14. This suggests that there is no evidence to reject the null hypothesis, stating that the two means are equal. However, the p-value is not very large, so this is not a 'strong evidence'. But this suggests that we should choose Racing games for their new channel.

Concerns about the assumptions

The Welch two sample t-test is designed for unequal population variances, which is strongly satisfied in our case, but assume normality for the two populations compared. .



As we can observe on the histograms above, the numbers of viewers for the two genres follow a right skewed distribution, whereas we expect a symmetric one for a normal population. Therefore, we may be worried about the validity of our assumptions and the meaningfulness of the hypothesis test conducted. As such, no definitive conclusion can be drawn from it.