Team Abolishes
Dawei Wang

In last milestone, a first order term of Video Games Sales in European was built, and it has a 0.4286 adjusted R^2. In this week, a further and deep refining process of this model was conducted.

To discover the if there's a significant interaction in my dataset, two new interaction terms was built: CriticScore_CriticCount and UserScore_UserCount. Considering there might be an unclear relationship between the score and counts from either critics or users, I made these two terms and try to make an improvement. The coefficients show below:

```
CriticScore_CriticCount  3.101e-04  6.670e-05   4.649 3.43e-06 ***
UserScore_UserCount      9.399e-06  4.308e-06   2.182 0.029184 *
```

With these two terms, the new model has slightly improvement with 0.439 adjusted R^2 and good F-test result.

Then all 4 numeric independent variables that produced by Critic_Count, CriticScore, UserScore and UserCount was added to the dataset respectively.

```
CriticC_Sq               3.676e-05  3.805e-05   0.966 0.334119
CriticS_Sq               4.019e-05  7.810e-05   0.515 0.606833
UserS_Sq                -1.804e-02  5.840e-03  -3.089 0.002021 **
UserC_Sq                 5.455e-09  4.271e-09   1.277 0.201524
---
```

However, their t-test doesn't show a reason to keep them stay in the model. By removing the biggest t-test term one by one, the $UserScore^2$ was the only second order term that is good enough to kept in the model.

Thus, the refined model become a model with two new interaction term and one new second order term, and has passed the F-test, T-test:

Log (EU_Sales) ~

PlatForm + Genre + Critic_Score + Critic_Count + User_Score + log (User_Count) + CriticScore_CriticCount + UserScore_UserCount + UserS_Sq

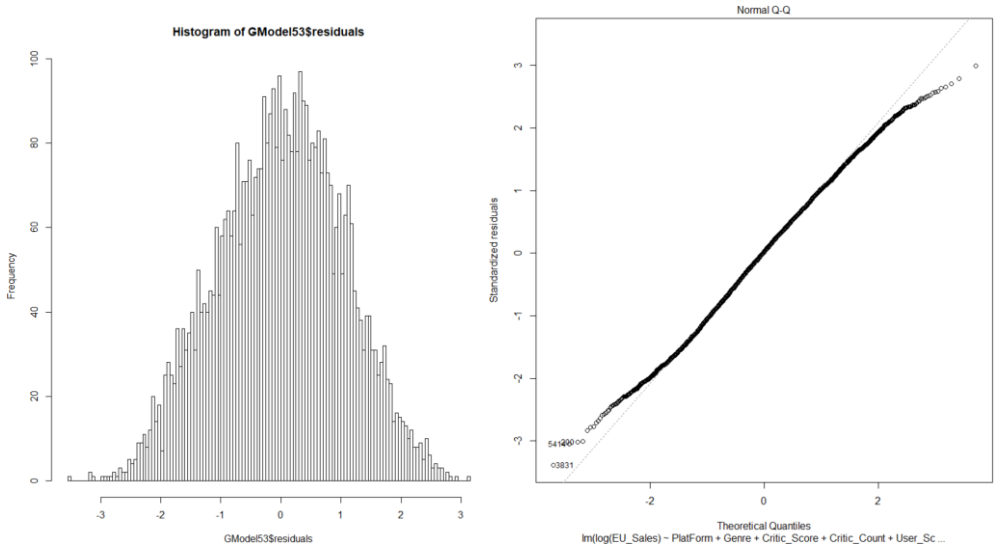
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -5.291e+00  2.538e-01 -20.843  < 2e-16 ***
PlatFormNintendo         3.788e-01  4.922e-02   7.697 1.69e-14 ***
PlatFormPC              -1.316e+00  6.570e-02 -20.024  < 2e-16 ***
PlatFormSony             5.337e-01  3.679e-02  14.508  < 2e-16 ***
GenreAdventure          -2.690e-01  8.691e-02  -3.095 0.001979 **
GenreFighting           -4.416e-02  7.065e-02  -0.625 0.532002
GenreMisc                4.058e-01  7.647e-02   5.307 1.17e-07 ***
GenrePlatform            1.996e-01  7.049e-02   2.831 0.004658 **
GenreRacing              3.398e-01  5.992e-02   5.671 1.51e-08 ***
GenreRole-Playing       -6.133e-01  5.748e-02 -10.670  < 2e-16 ***
GenreShooter            -4.061e-02  5.229e-02  -0.777 0.437475
GenreSimulation          3.699e-01  8.160e-02   4.533 5.97e-06 ***
GenreSports              1.843e-01  5.505e-02   3.347 0.000823 ***
GenreStrategy           -2.963e-01  8.440e-02  -3.511 0.000451 ***
Critic_Score             1.154e-02  2.297e-03   5.024 5.26e-07 ***
Critic_Count            -2.298e-02  5.233e-03  -4.391 1.16e-05 ***
User_Score               1.350e-01  7.236e-02   1.865 0.062202 .
log(User_Count)          4.780e-01  1.711e-02  27.933  < 2e-16 ***
CriticScore_CriticCount  3.378e-04  6.722e-05   5.025 5.23e-07 ***
UserScore_UserCount      9.681e-06  4.305e-06   2.249 0.024574 *
UserS_Sq                -1.734e-02  5.534e-03  -3.134 0.001737 **

Multiple R-squared:  0.4429,    Adjusted R-squared:  0.4405
F-statistic:   186 on 20 and 4680 DF,  p-value: < 2.2e-16
```

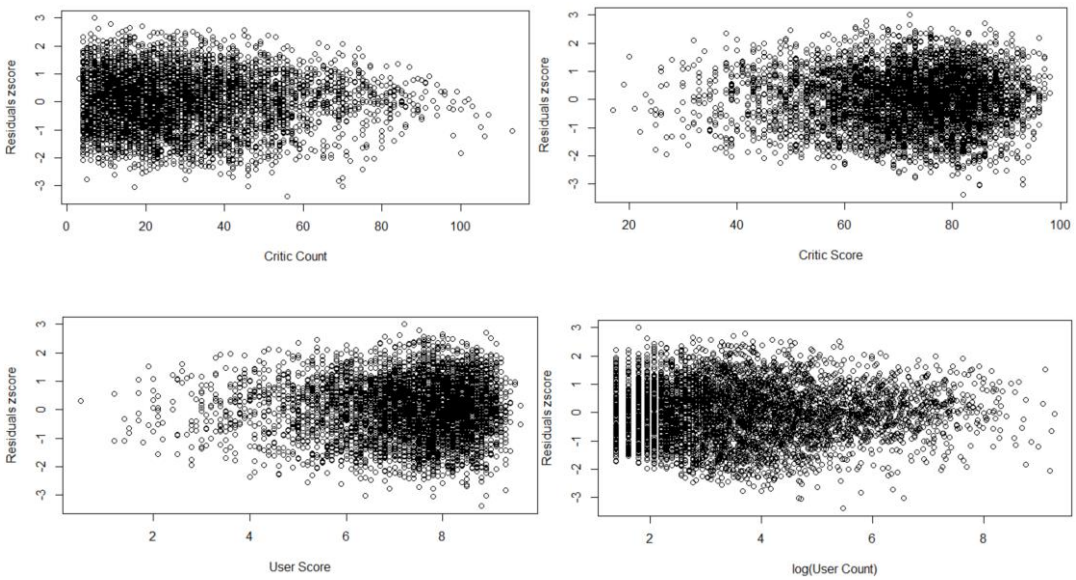
Here, the second order model UserScore^2 only brings a 0.0015 improvement on adjusted R^2. It may cost more resources to build the model or run the process if there's a lot more observations. It could be removed if needed. However, in this case, it didn't cost a lot more to get the conclusion, so, it was saved in the model.

All these steps above applied a training dataset contains 80% observations of original database and a testing dataset contains the left 20% observations of original database.

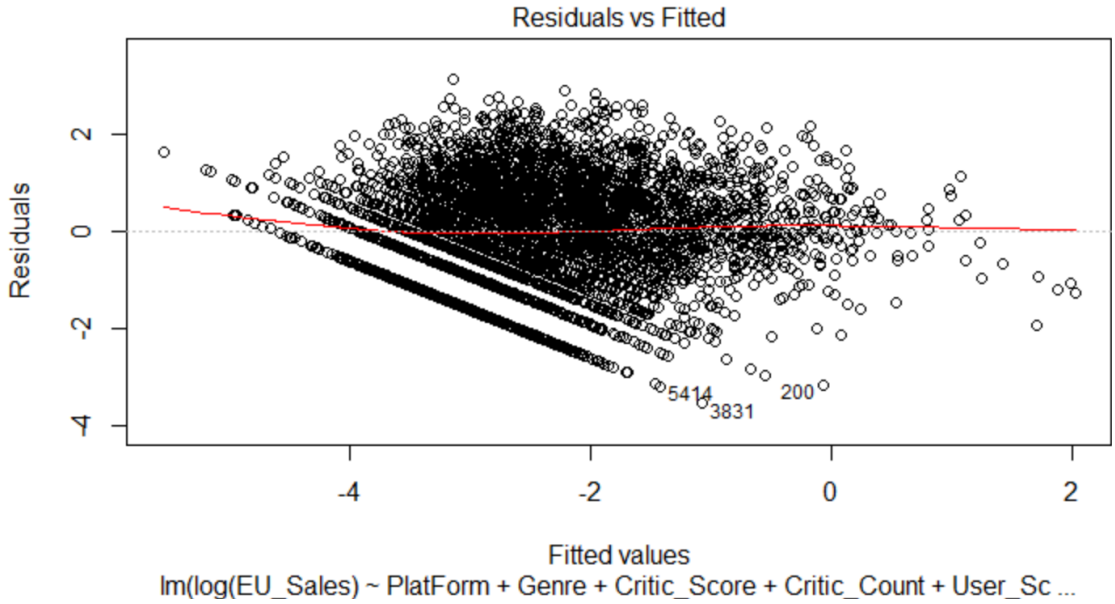
Residual Analysis



This is the histogram of the residual's distribution on the left. In this graph we can say, it's normally distributed. The right graph confirms residuals are normally distributed, though some of them on the both ends are a bit away from the y= x line.



The 4 graphs listed above show residuals vs the different variables. User count has been transformed here. In these graphs, we can not find some certain patterns in and no big change in variance of the residuals from one side to the next. In addition, most of data are distributed around 0 and about 95% sit within two standard deviations. It looks healthy.



In this graph we can see, the variances are mostly even and no big changes in

the middle mass. It's nearly homoscedastic and relatively healthy.

```
> durbinWatsonTest(GModel53)
 lag Autocorrelation D-W Statistic p-value
   1       0.2342245      1.530596       0
 Alternative hypothesis: rho != 0
```

The test score shows residuals' independency .

In sum, a regression model of European Video Game Sales was built. It passed the p-test, t-test and residual analysis with a adjusted R^2 is 0.4405.