Predicting Video Game Sales

STAT 204 Fall 2019 FINAL

Meltem Ozcan
Gulzina Kuttubekova

Data

Categorical Variables: Name, Platform, Genre, Publisher, Developer, Rating

Numerical Variables: Year, Global_Sales, NA_Sales, EU_Sales, JP_Sales, Other_Sales, Critic_Score, Critic_Count, User_Score, User_Count

6825 complete cases

16 variables + **7** new variables

New Categorical Variables:
Platform_Generation, Family_Platform, Platform_Company, Main_Developer, Developer_Country, Decade, Main_Publisher

New Numerical Variables:
Years_since_Release

*used *not used

Goals

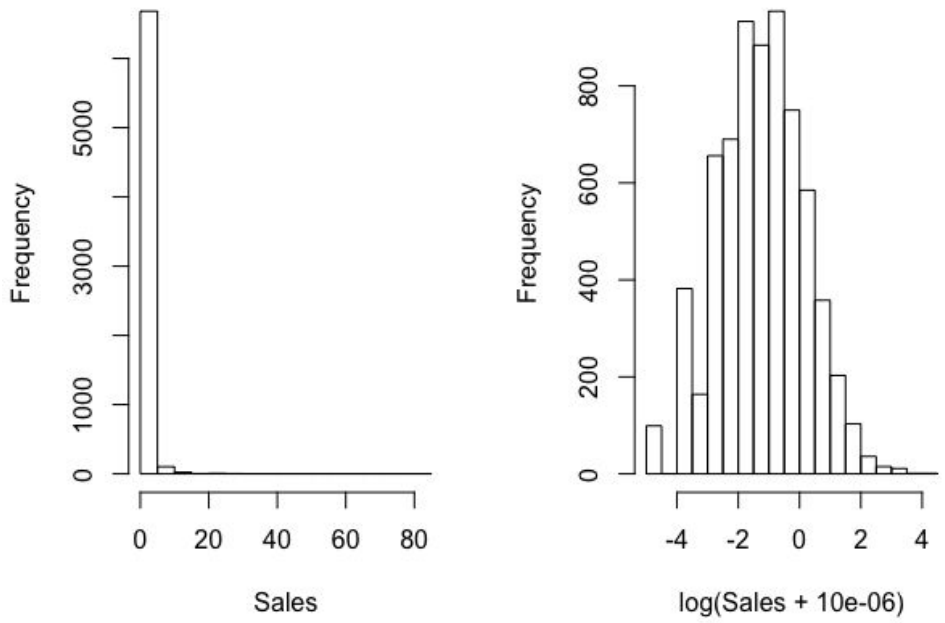
1) What properties of games best predict global sales?

2) Are the interactions* significant in predicting the global sales?

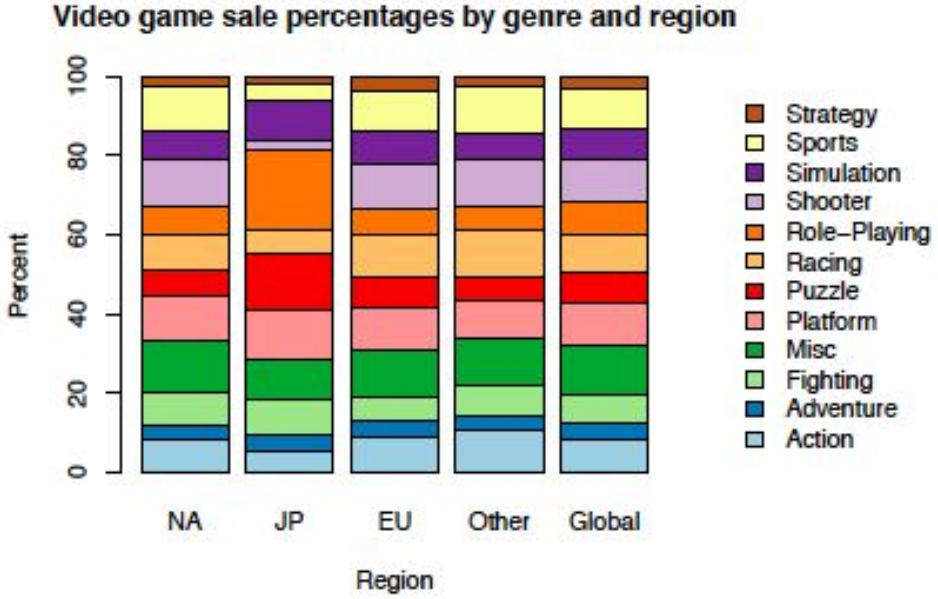
*We test different hypotheses related to the significance of interactions between pairs of categorical variables

Response variable - Sales

- Total sales in the world and disjoint regions (in millions of units)
- **Global Sales = NA + EU + JP + Other Sales**
- Use log(+epsilon)-transformation
  - Epsilon = 0.000001



1/4 Genre vs Region



$$y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ijk} \; \epsilon_{ijk} \sim N(0, \sigma)$$

where $k = 1, ..., n_{ij}$, $i = 1, ..., 12$ and $j = 1, 2, 3, 4$.

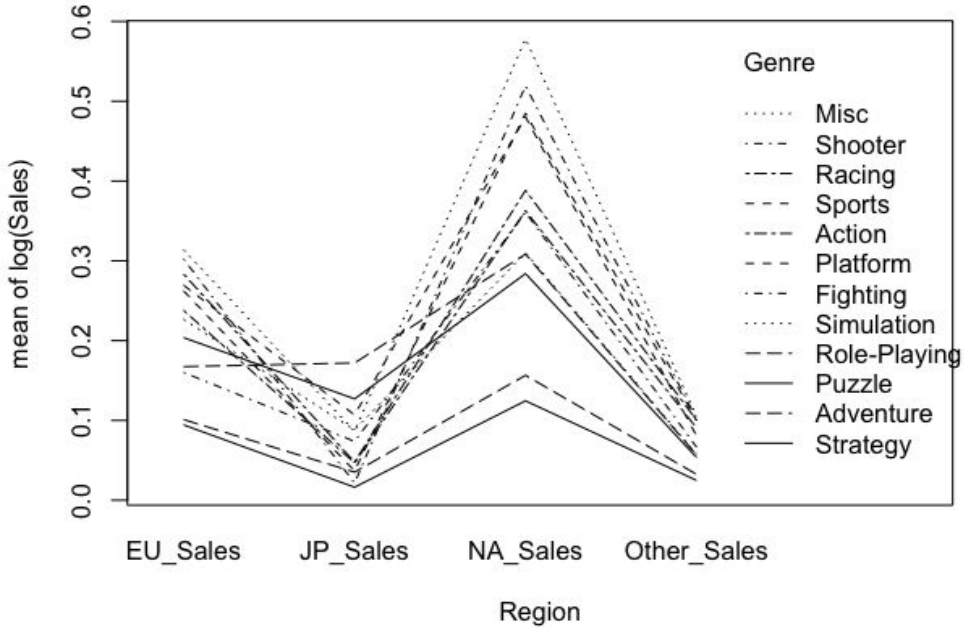
α_i is the main effect of genre type i

γ_j is the main effect of region type j

$H_0 : (\alpha\gamma)_{ij} = 0$ for all i, j

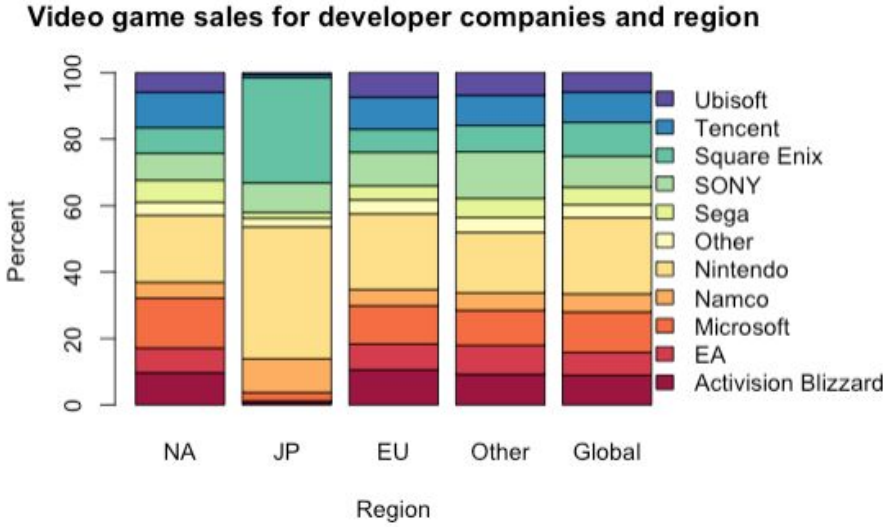
$H_1 : (\alpha\gamma)_{ij} \neq 0$ for at least one pair i, j

1/4 Genre vs Region




```
## Analysis of Variance Table
##
## Response: log(Sales)
##                Df Sum Sq Mean Sq  F value    Pr(>F)
## Genre          11   6586     599   34.032 < 2.2e-16 ***
## Region          3 243201   81067 4607.801 < 2.2e-16 ***
## Genre:Region   33  16967     514   29.224 < 2.2e-16 ***
## Residuals   27252 479456      18
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2/4 Developer Company vs Region



$$y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \epsilon_{ijk} \; \epsilon_{ijk} \sim N(0, \sigma)$$

where $k = 1, ..., n_{ij}$, $i = 1, ..., 11$ and $j = 1, 2, 3, 4$.

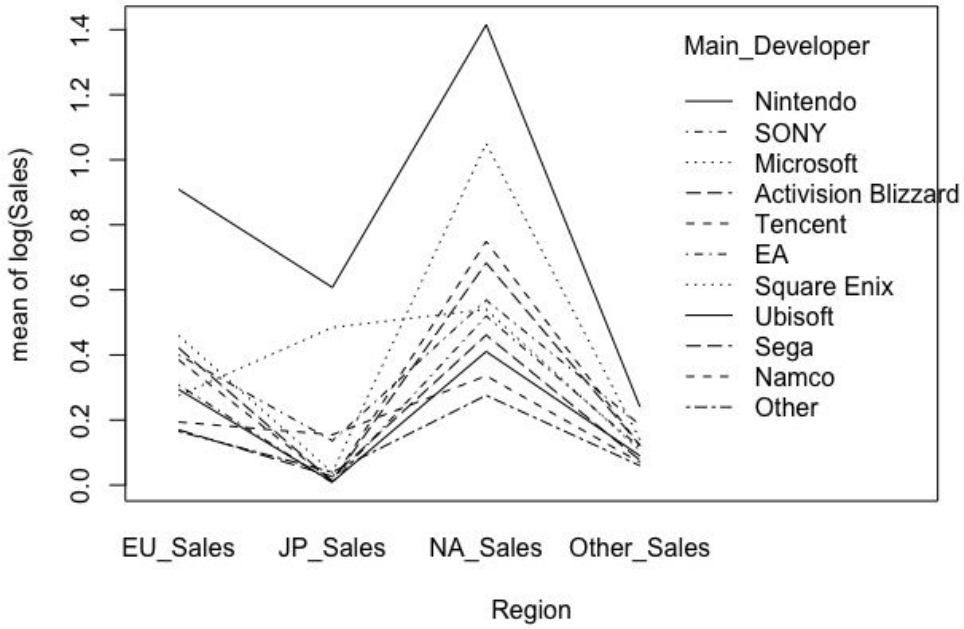
α_i is the main effect of developer type i

γ_j is the main effect of region type j

$H_0 : (\alpha\gamma)_{ij} = 0$ for all i, j

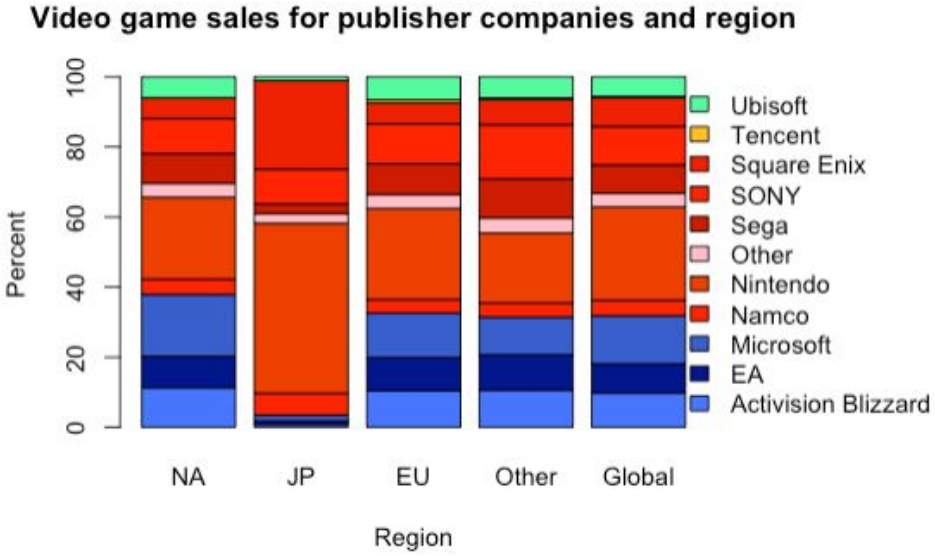
$H_1 : (\alpha\gamma)_{ij} \neq 0$ for at least one pair i, j

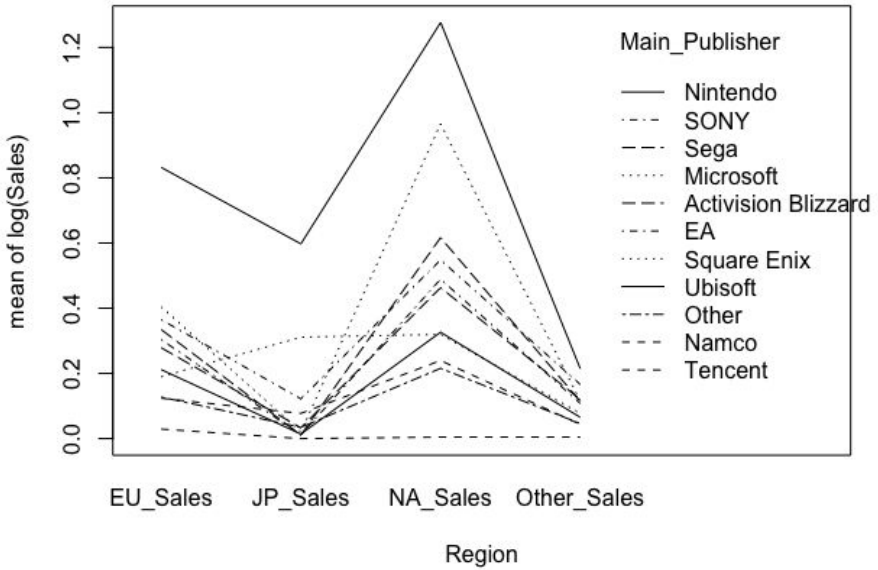
2/4 Developer Company vs Region




```
## Analysis of Variance Table
##
## Response: log(Sales)
##                          Df Sum Sq Mean Sq  F value    Pr(>F)
## Main_Developer           10  15298    1530   88.180 < 2.2e-16 ***
## Region                    3 243201   81067 4672.672 < 2.2e-16 ***
## Main_Developer:Region    30  14842     495   28.516 < 2.2e-16 ***
## Residuals             27256 472869      17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

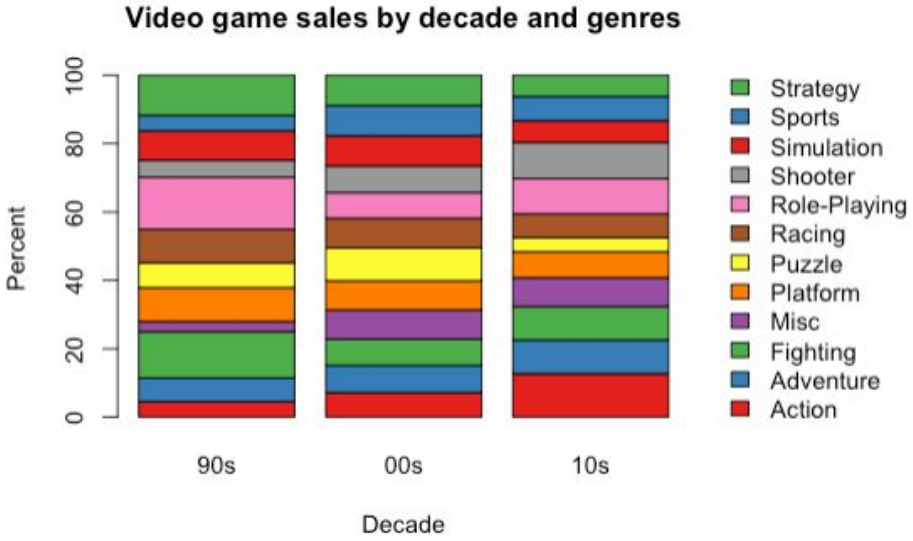
3/4 Publisher Company vs Region

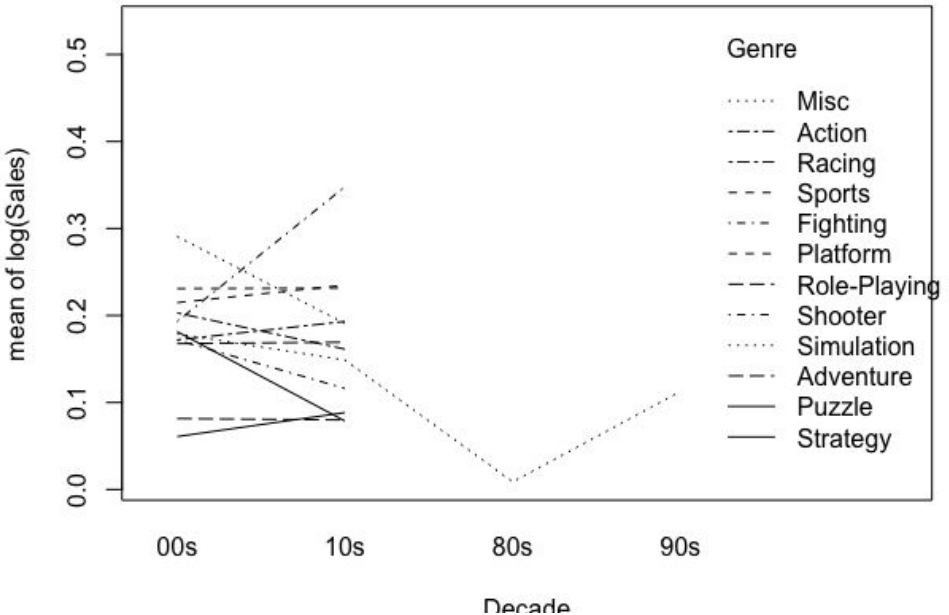





```
## Analysis of Variance Table
##
## Response: log(Sales)
##                          Df Sum Sq Mean Sq  F value    Pr(>F)
## Main_Publisher           10  22276    2228  132.319 < 2.2e-16 ***
## Region                    3 243201   81067 4815.478 < 2.2e-16 ***
## Main_Publisher:Region    30  21888     730   43.339 < 2.2e-16 ***
## Residuals             27256 458846      17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

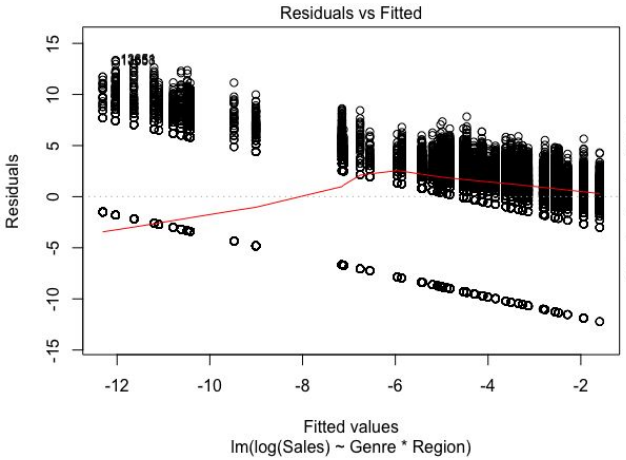
4/4 Genre vs Decade

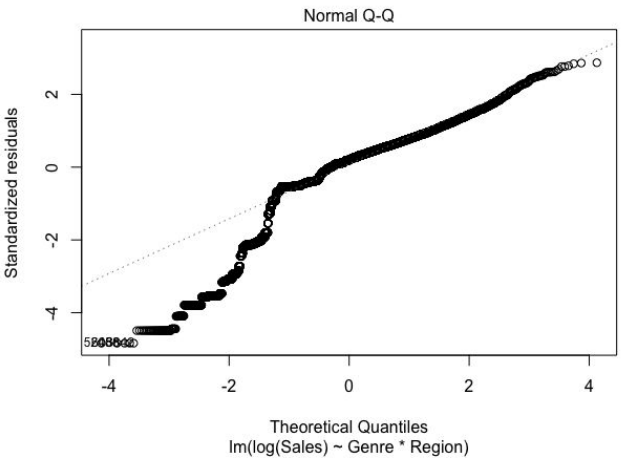


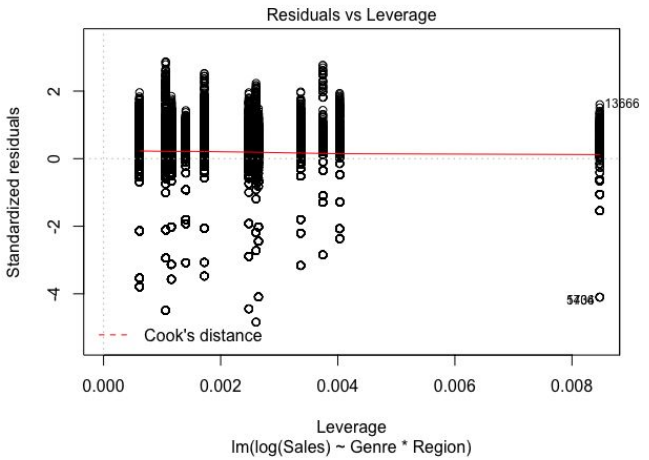



```
## Analysis of Variance Table
##
## Response: log(Sales)
##                  Df Sum Sq Mean Sq F value    Pr(>F)
## Decade            3   3709 1236.27 45.8978 < 2.2e-16 ***
## Genre            11   6043  549.32 20.3941 < 2.2e-16 ***
## Decade:Genre     22   2122   96.45  3.5808 2.655e-08 ***
## Residuals    27263 734337   26.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking assumptions for the 4 hypothesis tests







Checking Assumptions

	Method	Independence	Homogeneity (Breusch-Pagan test)	Normality (Shapiro-Wilk test)
Genre vs Region	LS	yes	p-value < 2.2e-16	p-value < 2.2e-16
	WLS	yes	p-value < 2.2e-16	p-value < 2.2e-16
Developer vs Region	LS	yes	p-value < 2.2e-16	p-value < 2.2e-16
	WLS	yes	p-value < 2.2e-16	p-value < 2.2e-16
Publisher vs Region	LS	yes	p-value < 2.2e-16	p-value < 2.2e-16
	WLS	yes	p-value < 2.2e-16	p-value < 2.2e-16
Genre vs Decade	LS	yes	p-value < 2.2e-16	p-value < 2.2e-16
	WLS	yes	p-value < 2.2e-16	p-value < 2.2e-16

Model 1: without interactions

- Split ‘sales.csv’ into training (70%) and test (30%) set
- Use **stratified partitioning** method to ensure each level of each categorical variable represented in equal proportion in each set
- Include all explanatory variables mentioned previously

```
## Response: log(Sales)
##                    Df Sum Sq Mean Sq  F value    Pr(>F)
## Genre              11   5619     511   34.684 < 2.2e-16 ***
## Critic_Score        1  18614   18614 1263.851 < 2.2e-16 ***
## Critic_Count        1  25503   25503 1731.613 < 2.2e-16 ***
## Rating              6   1017     169   11.507 6.719e-13 ***
## Decade              3   3151    1050   71.311 < 2.2e-16 ***
## Platform_Company    3  18013    6004  407.694 < 2.2e-16 ***
## Platform_Gen        3   9145    3048  206.984 < 2.2e-16 ***
## Family_Platform     3   4060    1353   91.898 < 2.2e-16 ***
## Main_Developer     10   4143     414   28.130 < 2.2e-16 ***
## Main_Publisher     10   2998     300   20.359 < 2.2e-16 ***
## Region              3 213781   71260 4838.447 < 2.2e-16 ***
## Residuals       24029 353897      15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2: with interactions

- Use same train dataset
- Include interactions found significant in the previous hypothesis tests

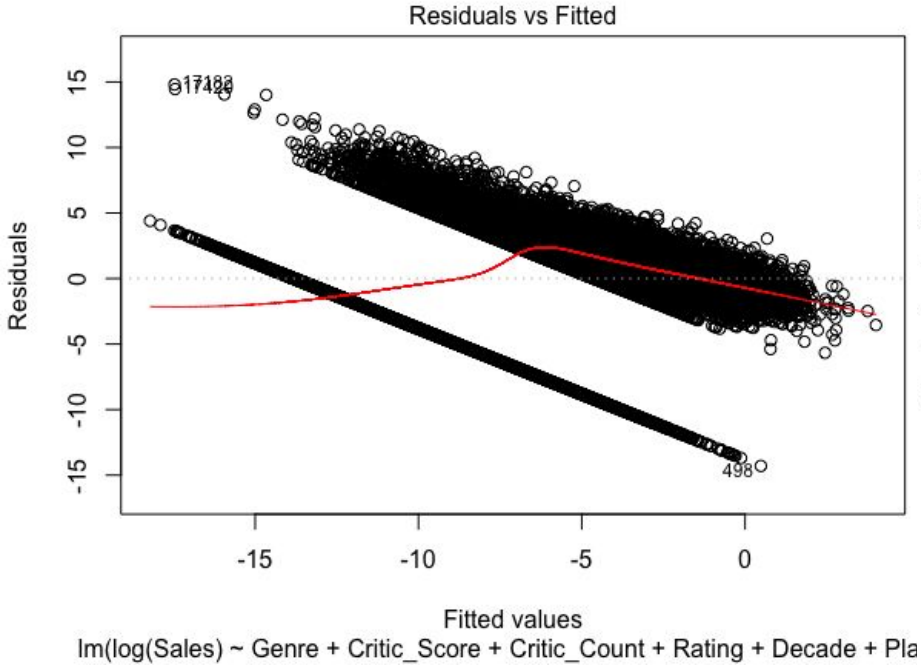
```
## Analysis of Variance Table
##
## Response: log(Sales)
##                          Df Sum Sq Mean Sq   F value    Pr(>F)
## Genre                    11   5619     511   37.8286 < 2.2e-16 ***
## Critic_Score              1  18614   18614 1378.4469 < 2.2e-16 ***
## Critic_Count              1  25503   25503 1888.6209 < 2.2e-16 ***
## Rating                    6   1017     169   12.5502 3.509e-14 ***
## Decade                    3   3151    1050   77.7772 < 2.2e-16 ***
## Platform_Company          3  18013    6004  444.6605 < 2.2e-16 ***
## Platform_Gen              3   9145    3048  225.7521 < 2.2e-16 ***
## Family_Platform           3   4060    1353  100.2304 < 2.2e-16 ***
## Main_Developer           10   4143     414   30.6805 < 2.2e-16 ***
## Main_Publisher           10   2998     300   22.2051 < 2.2e-16 ***
## Region                    3 213781   71260 5277.1569 < 2.2e-16 ***
## Genre:Region             33  14570     442   32.6968 < 2.2e-16 ***
## Main_Developer:Region    30   8209     274   20.2643 < 2.2e-16 ***
## Main_Publisher:Region    30   6551     218   16.1703 < 2.2e-16 ***
## Genre:Decade             22   1644      75    5.5327 9.463e-16 ***
## Residuals             23914 322923      14
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

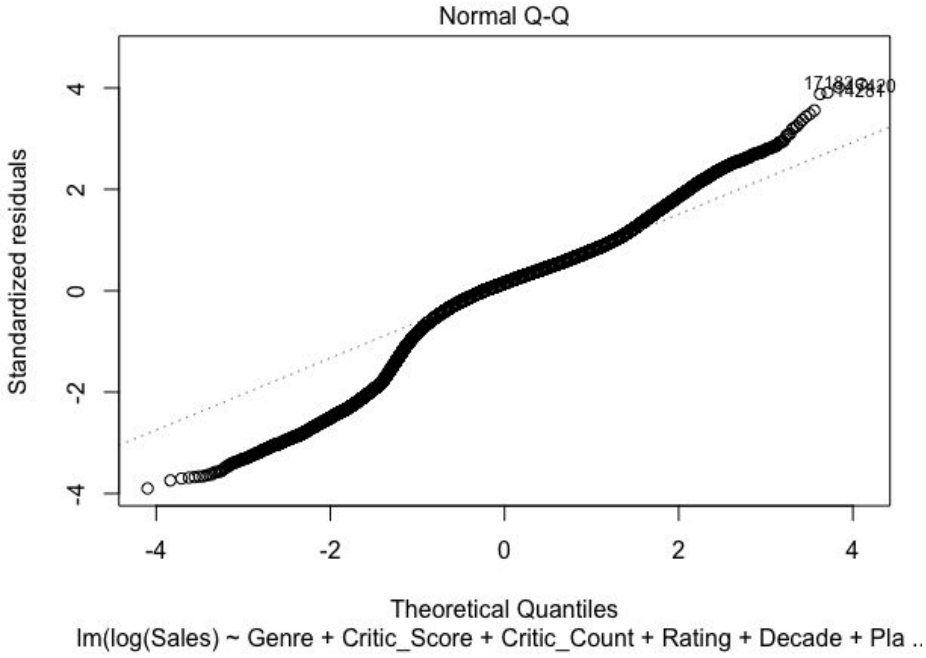
Comparing two models

Model	**AIC**	**BIC**	**Adj-R^2**	**train RMSE**	**test RMSE**
MLR without interactions	133184.3	133637.3	0.4625	6.774657	6.494338
MLR with interactions	131208.4	132591.7	0.5072	6.862823	6.650431

There is an odd relationship between train RMSE and test RMSE, potentially caused by underfitting. This can be explained by the famous ”curse of dimensionality” and variance-bias trade-off. For instance, we have a high bias and relatively small variance, which leads to these results.

Checking Assumptions - (Both models)





Checking Assumptions - (Both models)

What we did:

- Addressed multicollinearity by dropping some numeric variables
- Dropped highly dependent categorical variables by EDA and inductive bias
- Log-transformed the response variable

What we got:

- There is an obvious violation of homogeneity of variance (Breusch-Pagan p-value < 2.2e-16)
- Normality assumption is violated (seen in both QQ and residual plots. Shapiro-Wilk p-value < 2.2e-16)
- No outliers, consequently no influential points observed

Conclusions + Future Directions

- How we dealt with challenges?
  - Different transformations on response variable (log, sqrt, etc.)
  - EDA for detecting multicollinearity and dependence
  - LS and WLS methods
- Faced new challenges:
  - Bias-variance trade-off
  - Curse of dimensionality
- Room for improvement:
  - Collect more data
  - Reduce number of parameters, by selecting only few explanatory variables
  - Transform coefficient parameters

References

https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings

https://www.vgchartz.com/analysis/platform_totals/

https://en.wikipedia.org/wiki/Video_game_console