The predicting model

Building the model

Supposed there is a new programmer want to enter Google Play market, he/she have some ideas in some Genres, can hire professional tester for initial rating and can make Free app with ads or Paid app. All he/she cares is the number of Installs, the higher number the more profit he get.

The selected data as below

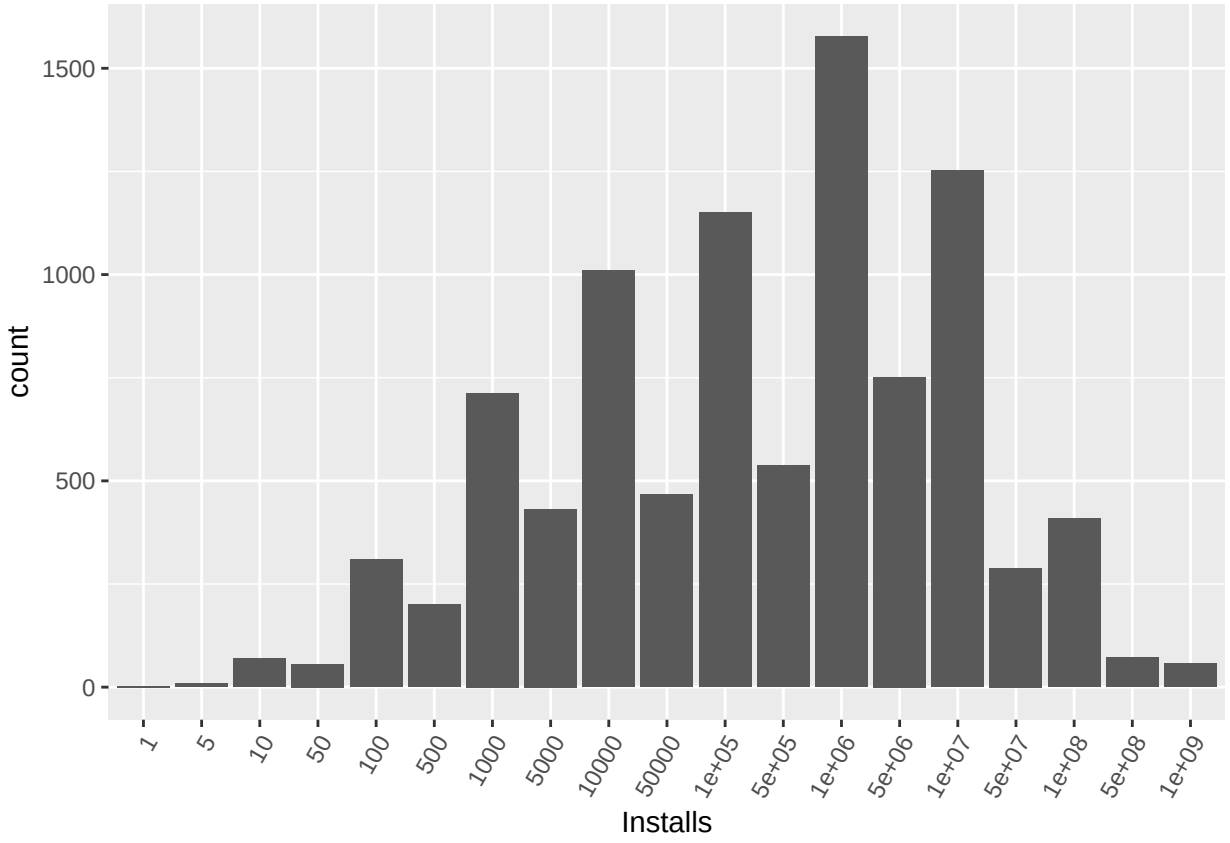
```
head(df_selected)
```

```
##   Rating    Installs Type       Genres
## 1    4.1     10,000+ Free Art & Design
## 2    3.9    500,000+ Free Art & Design
## 3    4.7  5,000,000+ Free Art & Design
## 4    4.5 50,000,000+ Free Art & Design
## 5    4.3    100,000+ Free Art & Design
## 6    4.4     50,000+ Free Art & Design
```

```
summary(df_selected)
```

```
##      Rating             Installs      Type                Genres
##  Min.   :1.000   1,000,000+ :1577   Free:8719   Tools        : 734
##  1st Qu.:4.000   10,000,000+:1252   Paid: 647   Entertainment: 577
##  Median :4.300   100,000+   :1150               Education    : 563
##  Mean   :4.192   10,000+    :1010               Action       : 375
##  3rd Qu.:4.500   5,000,000+ : 752               Productivity : 351
##  Max.   :5.000   1,000+     : 713               Medical      : 350
##                  (Other)    :2912               (Other)      :6416
```

The distribution of current number of Installs



Ordinal Logistic Regression model

Since the number of Installs in this dataset is classified as categorial value, Ordinal Logistic Regression seems to be the most suitable.

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

##
## Re-fitting to get Hessian

## Call:
## polr(formula = Installs ~ Rating + Type + Genres, data = df_selected)
##
## Coefficients:
##                                Value Std. Error  t value
## Rating                       0.63020    0.03670  17.1703
## TypePaid                    -2.04403    0.07284 -28.0617
## GenresAdventure             -0.52623    0.20702  -2.5419
## GenresArcade                 0.11741    0.15037   0.7808
## GenresArt & Design          -1.66303    0.21746  -7.6476
## GenresAuto & Vehicles       -1.84076    0.21119  -8.7160
```

```
## GenresBeauty                  -1.90679    0.26408  -7.2205
## GenresBoard                   -1.22506    0.23483  -5.2169
## GenresBooks & Reference       -1.65182    0.15892 -10.3941
## GenresBusiness                -1.93254    0.13882 -13.9213
## GenresCard                    -0.92597    0.25618  -3.6146
## GenresCasino                  -1.16710    0.28515  -4.0930
## GenresCasual                  -0.22297    0.14281  -1.5614
## GenresComics                  -1.61977    0.22966  -7.0529
## GenresCommunication           -0.03418    0.13941  -0.2452
## GenresDating                  -1.51624    0.15214  -9.9662
## GenresEducation               -2.02832    0.11715 -17.3144
## GenresEducational             -0.98654    0.18325  -5.3837
## GenresEntertainment           -1.25552    0.11709 -10.7224
## GenresEvents                  -2.76719    0.28126  -9.8386
## GenresFinance                 -1.65381    0.13049 -12.6741
## GenresFood & Drink            -0.85206    0.17840  -4.7762
## GenresHealth & Fitness        -0.98907    0.13128  -7.5341
## GenresHouse & Home            -0.87161    0.20017  -4.3544
## GenresLibraries & Demo        -1.98518    0.22284  -8.9087
## GenresLifestyle               -1.77203    0.13447 -13.1777
## GenresMaps & Navigation       -1.19447    0.17997  -6.6371
## GenresMedical                 -2.25667    0.12898 -17.4966
## GenresMusic                   -0.40867    0.36228  -1.1280
## GenresMusic & Audio           -1.24591    1.42060  -0.8770
## GenresNews & Magazines        -1.23907    0.14689  -8.4356
## GenresParenting               -1.69598    0.24172  -7.0163
## GenresPersonalization         -1.30585    0.13602  -9.6003
## GenresPhotography              0.04566    0.13338   0.3423
## GenresProductivity            -0.64664    0.13439  -4.8118
## GenresPuzzle                  -0.35461    0.16672  -2.1271
## GenresRacing                   0.09348    0.18265   0.5118
## GenresRole Playing            -0.52781    0.17243  -3.0611
## GenresShopping                -0.31274    0.14324  -2.1834
## GenresSimulation              -0.84702    0.14503  -5.8404
## GenresSocial                  -0.52044    0.14687  -3.5436
## GenresSports                  -0.74480    0.12966  -5.7442
## GenresStrategy                -0.03169    0.18440  -0.1719
## GenresTools                   -1.24214    0.11175 -11.1156
## GenresTravel & Local          -0.67842    0.14709  -4.6122
## GenresTrivia                  -1.78844    0.32816  -5.4498
## GenresVideo Players & Editors -0.50828    0.16647  -3.0533
## GenresWeather                 -0.54695    0.20562  -2.6600
## GenresWord                    -0.30701    0.33510  -0.9162
##
## Intercepts:
##              Value    Std. Error t value
## 1|5           -6.8906   0.5991   -11.5017
## 5|10          -5.7092   0.3375   -16.9184
## 10|50         -3.7950   0.2076   -18.2806
## 50|100        -3.2491   0.1952   -16.6435
## 100|500       -1.9928   0.1824   -10.9286
## 500|1000      -1.5775   0.1806    -8.7367
## 1000|5000     -0.6640   0.1788    -3.7128
## 5000|10000    -0.2838   0.1786    -1.5886
```

```
## 10000|50000   0.4123   0.1788     2.3059
## 50000|1e+05   0.6867   0.1790     3.8353
## 1e+05|5e+05   1.3084   0.1797     7.2803
## 5e+05|1e+06   1.5880   0.1801     8.8185
## 1e+06|5e+06   2.4394   0.1811    13.4720
## 5e+06|1e+07   2.9189   0.1817    16.0655
## 1e+07|5e+07   4.1060   0.1841    22.3074
## 5e+07|1e+08   4.5970   0.1859    24.7220
## 1e+08|5e+08   6.1025   0.2013    30.3204
## 5e+08|1e+09   6.9229   0.2238    30.9359
##
## Residual Deviance: 44511.82
## AIC: 44645.82
```

```
##
## Re-fitting to get Hessian
```

Calculating p value and filtering out those who have p value > 0.05 or have impact on the model

```
summary_table_filtered <- as_data_frame(summary_table, rownames = 'id')
summary_table_filtered <- summary_table_filtered %>%
                          filter(`p value` <= 0.05)
print.data.frame(summary_table_filtered)
```

```
##                             id      Value Std. Error    t value p value
## 1                       Rating  0.6301979 0.03670285  17.170270   0.000
## 2                     TypePaid -2.0440308 0.07284049 -28.061736   0.000
## 3               GenresAdventure -0.5262343 0.20702127  -2.541933   0.011
## 4           GenresArt & Design -1.6630345 0.21745828  -7.647603   0.000
## 5        GenresAuto & Vehicles -1.8407585 0.21119403  -8.715959   0.000
## 6                 GenresBeauty -1.9067931 0.26408221  -7.220453   0.000
## 7                  GenresBoard -1.2250628 0.23482586  -5.216899   0.000
## 8      GenresBooks & Reference -1.6518217 0.15891898 -10.394112   0.000
## 9              GenresBusiness -1.9325370 0.13881916 -13.921256   0.000
## 10                GenresCard -0.9259674 0.25617721  -3.614558   0.000
## 11              GenresCasino -1.1671018 0.28514593  -4.092998   0.000
## 12              GenresComics -1.6197745 0.22966119  -7.052887   0.000
## 13              GenresDating -1.5162436 0.15213911  -9.966166   0.000
## 14           GenresEducation -2.0283242 0.11714636 -17.314445   0.000
## 15         GenresEducational -0.9865431 0.18324699  -5.383680   0.000
## 16       GenresEntertainment -1.2555164 0.11709312 -10.722376   0.000
## 17              GenresEvents -2.7671917 0.28125812  -9.838620   0.000
## 18             GenresFinance -1.6538129 0.13048811 -12.674051   0.000
## 19         GenresFood & Drink -0.8520641 0.17839914  -4.776167   0.000
## 20      GenresHealth & Fitness -0.9890689 0.13127816  -7.534147   0.000
## 21         GenresHouse & Home -0.8716068 0.20016581  -4.354424   0.000
## 22      GenresLibraries & Demo -1.9851843 0.22283594  -8.908726   0.000
## 23            GenresLifestyle -1.7720274 0.13447124 -13.177742   0.000
## 24     GenresMaps & Navigation -1.1944692 0.17996913  -6.637079   0.000
## 25              GenresMedical -2.2566736 0.12897766 -17.496624   0.000
## 26      GenresNews & Magazines -1.2390675 0.14688591  -8.435577   0.000
## 27            GenresParenting -1.6959820 0.24171890  -7.016340   0.000
## 28       GenresPersonalization -1.3058521 0.13602259  -9.600259   0.000
## 29         GenresProductivity -0.6466377 0.13438695  -4.811760   0.000
## 30              GenresPuzzle -0.3546140 0.16671601  -2.127054   0.033
## 31         GenresRole Playing -0.5278133 0.17242503  -3.061118   0.002
```

```
## 32                GenresShopping -0.3127373 0.14323655  -2.183362   0.029
## 33              GenresSimulation -0.8470201 0.14502779  -5.840398   0.000
## 34                  GenresSocial -0.5204433 0.14686912  -3.543586   0.000
## 35                  GenresSports -0.7447952 0.12966127  -5.744161   0.000
## 36                   GenresTools -1.2421393 0.11174777 -11.115563   0.000
## 37          GenresTravel & Local -0.6784180 0.14709085  -4.612238   0.000
## 38                  GenresTrivia -1.7884377 0.32816484  -5.449815   0.000
## 39 GenresVideo Players & Editors -0.5082763 0.16647026  -3.053256   0.002
## 40                  GenresWeather -0.5469475 0.20562085  -2.659981   0.008
## 41                            1|5 -6.8905709 0.59909192 -11.501692   0.000
## 42                           5|10 -5.7091829 0.33745433 -16.918387   0.000
## 43                          10|50 -3.7949690 0.20759553 -18.280591   0.000
## 44                         50|100 -3.2491192 0.19521856 -16.643495   0.000
## 45                        100|500 -1.9928335 0.18235060 -10.928582   0.000
## 46                       500|1000 -1.5774849 0.18055832  -8.736706   0.000
## 47                      1000|5000 -0.6639884 0.17883660  -3.712822   0.000
## 48                    10000|50000  0.4123486 0.17882398   2.305891   0.021
## 49                    50000|1e+05  0.6866579 0.17903744   3.835276   0.000
## 50                    1e+05|5e+05  1.3083908 0.17971591   7.280328   0.000
## 51                    5e+05|1e+06  1.5879878 0.18007384   8.818537   0.000
## 52                    1e+06|5e+06  2.4394254 0.18107398  13.471982   0.000
## 53                    5e+06|1e+07  2.9189068 0.18168752  16.065533   0.000
## 54                    1e+07|5e+07  4.1059982 0.18406477  22.307356   0.000
## 55                    5e+07|1e+08  4.5970100 0.18594809  24.722008   0.000
## 56                    1e+08|5e+08  6.1025263 0.20126793  30.320411   0.000
## 57                    5e+08|1e+09  6.9228867 0.22378171  30.935891   0.000
```

Explaining the model

$$logit\ [P(Y \le j)] = \alpha_j - \sum \beta_i X_i$$

where $j = 1,\ldots,J-1$ and $i = 1,\ldots,M$

The basic of proportional odds model have mathematical fomulation:

With 'J' is sum of number of factors in number of Installs (J=18) and 'M' is total number of independent variables (M=3).

'j' is each factor in number of Installs, meanwhile 'i' is each independent variables, simply put: * i =1 refers to Rating

- i = 2 refers to Type
- i = 3 refers to Genres

Interpretation:

Comments on Coefficients: Only rating of the app have positive effect on number of installs, if the app is paid or belong to these genres below will have negative impact on its number of Installs.

Comments on intercept: take 1|5 as example: the odd of log that the app will have only 1 person installs the app versus the odd of log many people (>1) try the app

Apply the model to our case

Suppose again our programer have finished 2 apps with charateristic like belows:

App 1 have Rating 4 in Genres Educational and a Free app with ads

```
new_app <- data.frame('Rating'=4,'Type'='Free','Genres'='Educational')
round(predict(model_fit,new_app,type = "p"), 3)
```

```
##     1     5    10    50   100   500  1000  5000 10000 50000 1e+05 5e+05
## 0.000 0.000 0.004 0.003 0.020 0.014 0.057 0.040 0.106 0.054 0.144 0.070
## 1e+06 5e+06 1e+07 5e+07 1e+08 5e+08 1e+09
## 0.199 0.088 0.129 0.026 0.034 0.006 0.005
```

App 2 is rated at 3.5 and is Racing game with a price.

```
new_app_2 <- data.frame('Rating'=3.5,'Type'='Paid','Genres'='Racing')
round(predict(model_fit,new_app_2,type = "p"), 3)
```

```
##     1     5    10    50   100   500  1000  5000 10000 50000 1e+05 5e+05
## 0.001 0.002 0.015 0.012 0.066 0.042 0.147 0.083 0.171 0.067 0.135 0.050
## 1e+06 5e+06 1e+07 5e+07 1e+08 5e+08 1e+09
## 0.108 0.036 0.044 0.008 0.010 0.002 0.001
```

Results: The first app may have 20% (highest chance) to get 1 million downloads while the second app has 17% (highest) to get at least 10 thounsand downloads.

If he/she knows the cost for developing the apps, then he can setup ads rate and setting price level using Expected Value calculation.