684 mideterm project

Yaqi Huang

2017/12/19

Introduction

Since last time, I was trying to buy some wines to celebrate the birthday of a friend of mine, I realized that I am definitely not an expert in knowledge related to wines. I was impressed by reading the decription written down. Also, seemes to be a common sense that the higher the rate the wine has been graded, the higher the price it would be. But that is not always the case. The country it origined, the genre etc, could all be the important factors that would affect the price of the wine.

Then I found this dataset, which relates to wine and contains the following fields:

-Points: the number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80) -Title: the title of the wine review, which often contains the vintage if you're interested in extracting that feature -Variety: the type of grapes used to make the wine (ie Pinot Noir) -Description: a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc. -Country: the country that the wine is from -Province: the province or state that the wine is from -Region 1: the wine growing area in a province or state (ie Napa) -Region 2: sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank -Winery: the winery that made the wine -Designation: the vineyard within the winery where the grapes that made the wine are from -Price: the cost for a bottle of the wine -Taster Name: name of the person who tasted and reviewed the wine -Taster Twitter Handle: Twitter handle for the person who tasted and reviewed the wine

The variables in this dataset that I am interested in and will use for the further analysis are followings:

- Points
- Variety
- Description
- Country
- Province
- Price

Because the dataset is enormous and for variable “Variety”, there are so many categories involved, therefore, I limited the data to the top20 most reviewed varieties during the data clean process.

I have posted questions for myself and will be trying to answer by different methods, EDA and regressions.

For the EDA session, I would try to figure out:

- What is the distribution of prices awarded for the most reviewed varieties?
- Is there any relationship appeared between the points and price of the wine?
- What are the most often used words for description of the wine?

To do the above, I would produce several types of EDA and wordcould.

For the regression session:

- What is the best possible regression to estimate the price of the wine from the given variables?

To answer this question, I would try to fit lots of models, include linear models and multilevel regressions.

Data Clean and Organize

```
##                      X                country            description
##                 120975                     43                 111567
##            designation                 points                  price
##                  35777                     21                    390
##               province              region_1               region_2
##                    423                   1205                     18
##            taster_name taster_twitter_handle                  title
##                     20                     16                 110638
##                variety                winery
##                    698                 15855
```

Some overview of this dataset that I am interested in:

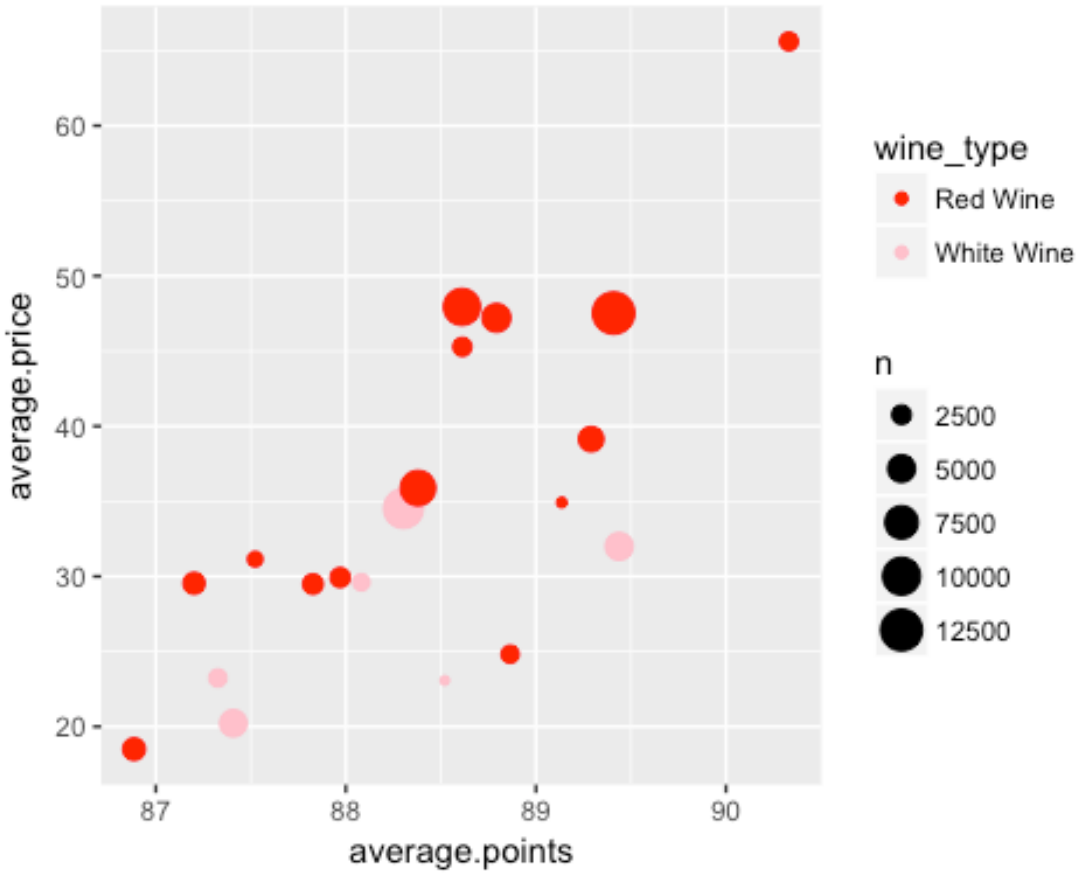
- The wines listed in the dataset come from 43 unique countries and 423 provinces.
- There are 698 varieties presented in the dataset.

As displayed above, the number of variety types are enormous, for the EDA session, I would be only focused on the top20 most reviewed variety as mentioned.

EDA

```
## # A tibble: 20 x 2
##                      variety count
##                       <fctr> <int>
```

```
##  1               Pinot Noir 12787
##  2               Chardonnay 11080
##  3       Cabernet Sauvignon  9386
##  4                Red Blend  8476
##  5 Bordeaux-style Red Blend  5340
##  6                 Riesling  4972
##  7          Sauvignon Blanc  4783
##  8                   Syrah  4086
##  9                   Rosé  3262
## 10                  Merlot  3062
## 11                Zinfandel  2708
## 12                  Malbec  2593
## 13               Sangiovese  2377
## 14                 Nebbiolo  2331
## 15           Portuguese Red  2196
## 16              White Blend  2172
## 17          Sparkling Blend  2027
## 18              Tempranillo  1789
## 19    Rhône-style Red Blend  1405
## 20               Pinot Gris  1391
```

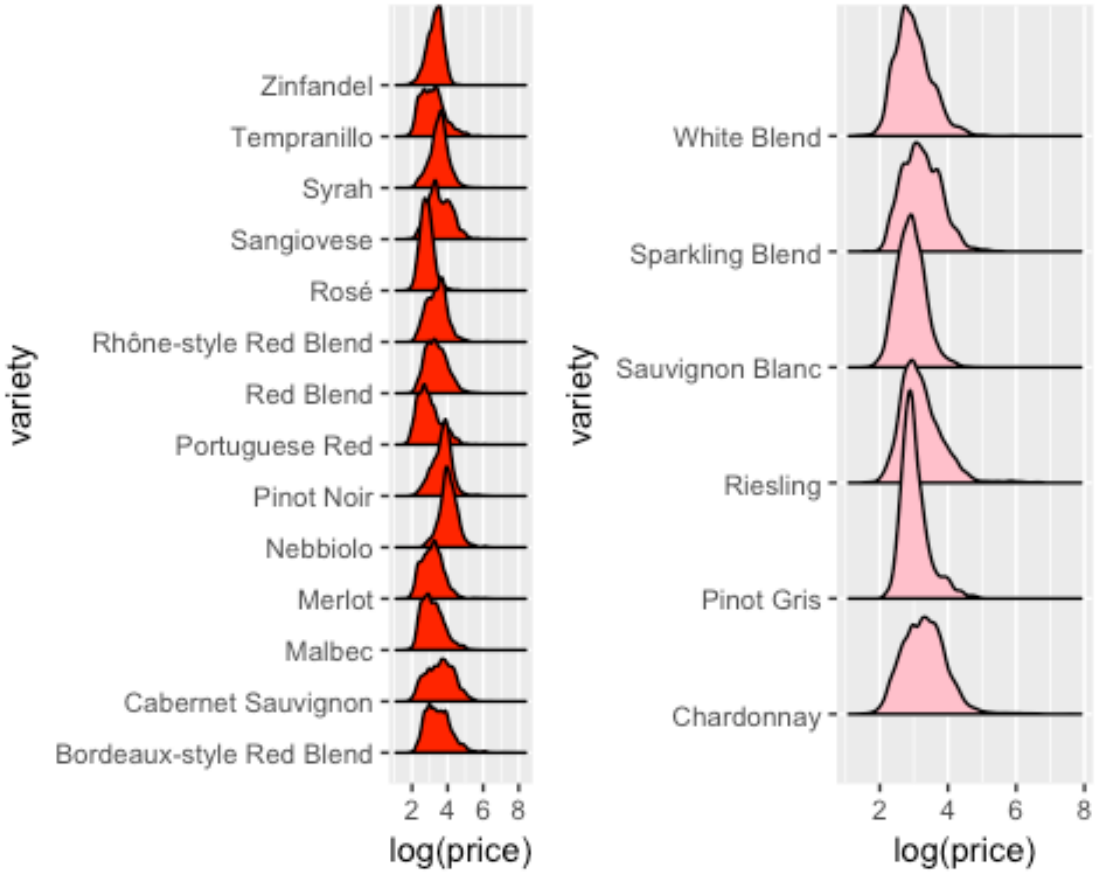


Something to be noted here that, after limit the dataset to the top20 most varieties reviewed, I created a new column named “wine_type”, which I put the top20 most

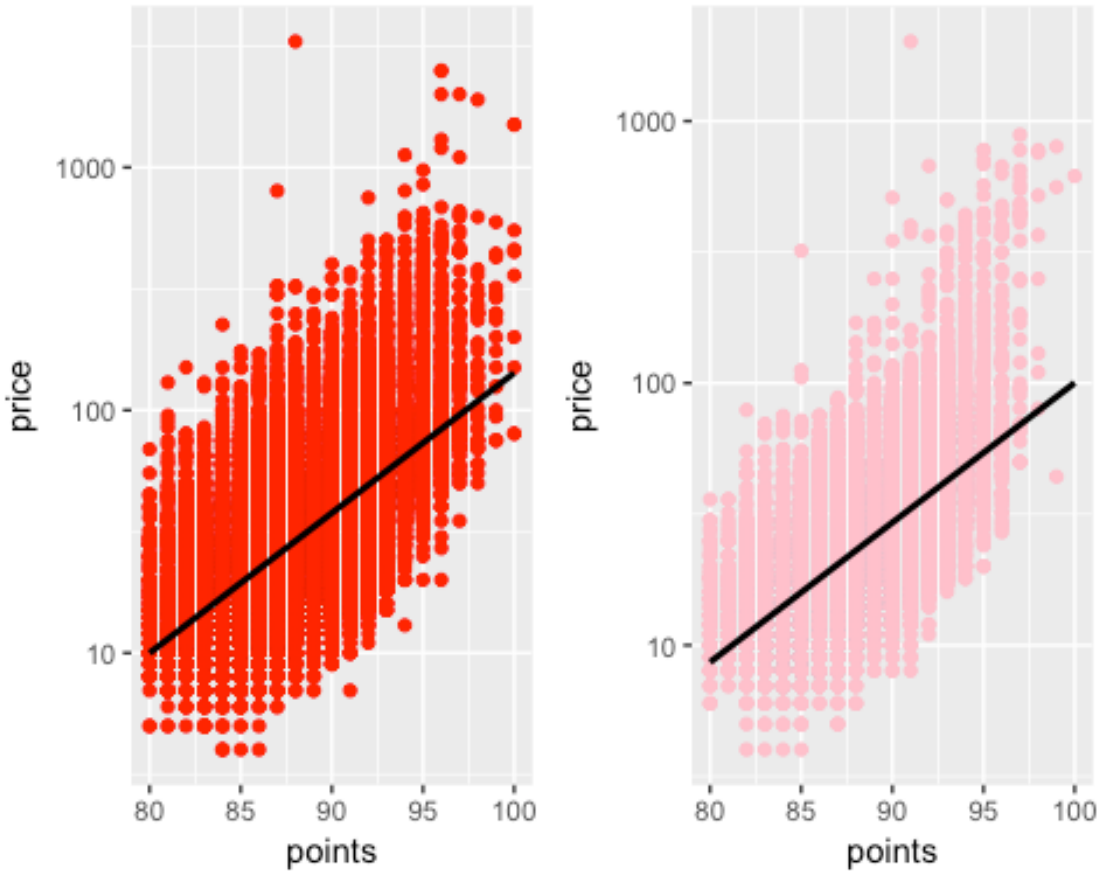
varieties viewed into either “Red Wine” or “White Wine”. To be noticed here, I put “Rosé” under “Red Wine”.

The top20 most varieties reviewed are displayed as table above and the correponding count numbers.

The plot above shown the average points against average price for the top20 most varieties reviewed, and color represents the different wine type that I classified, and the bigger the dot, the larger the corresponding count.



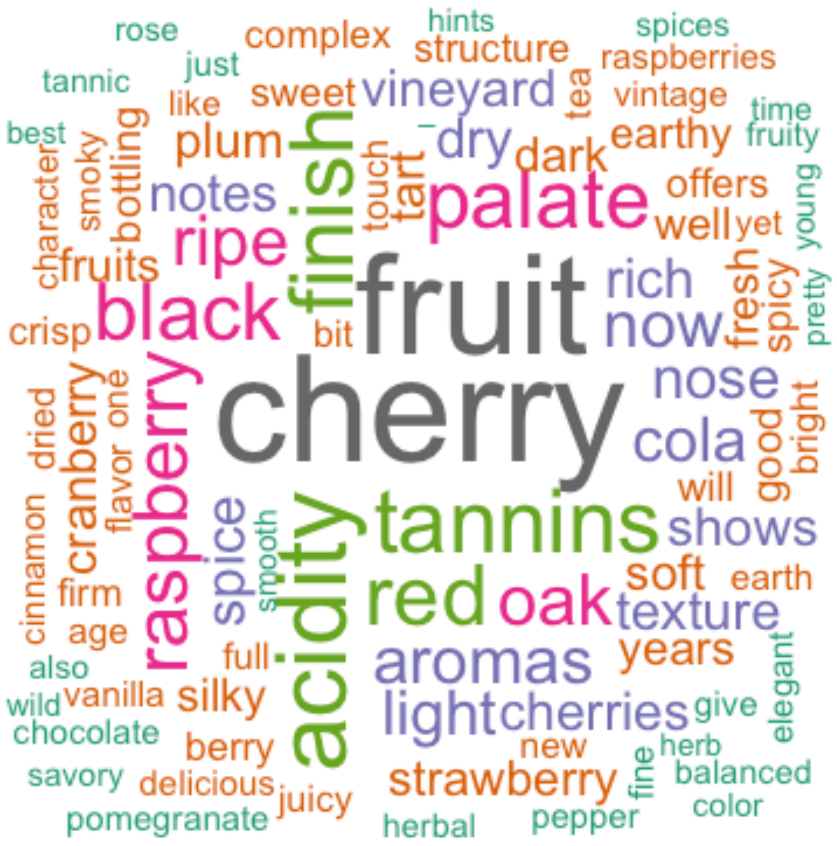
The above two plots could be used to answer the question of the distribution of the top20 most varieties reviewed. We could clear see the peak and the tails because of the closed polygon shape, but I do not think it clearly follows a normal distribution. Another point to mention from the above plots is that although there is some variation occured in the mean price for the red wine, but not big variation for the white wine. But possible because that the most varieties in top20 are categorized as red wine.



The above plots illustarte whether there is a relationship between the points awarded and the price of a given wine. The answer is yes, not to be so surprised. We could clearly see there is a positive relationship between the points awarded and the price, also pointed out by the trend line. But to be precise, what is the corresponding change for the price for a given increase in one unit for point, we could examine this in the regression below.

```
##                word freq
## cherry       cherry 5416
## fruit         fruit 5010
## acidity     acidity 3335
## tannins     tannins 2939
## finish       finish 2922
## red             red 2740
## palate       palate 2664
## black         black 2571
## raspberry raspberry 2408
## oak             oak 2182

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : concentrated could not be fit on page. It will not be plott
ed.
```

```
## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : complexity could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : aging could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : fullbodied could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : balance could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : along could not be fit on page. It will not be plotted.

## Warning in wordcloud(words = d$word, freq = d$freq, min.freq = 1, ma
x.words
## = 100, : showing could not be fit on page. It will not be plotted.
```



As from above, we got that the top most variety viewed is Pinot Noir in this dataset, therefor I am interested in that what kind of words have been written on the description in common, it could be a possible reason for a better reviewed as well.

From the wordcloud we could see that, most of the words used could cheer people up and are on the positive side of the language. Also there is a great number of fruits names used as well, which could possibly give the buyers a feeling of being natural when reading the description.

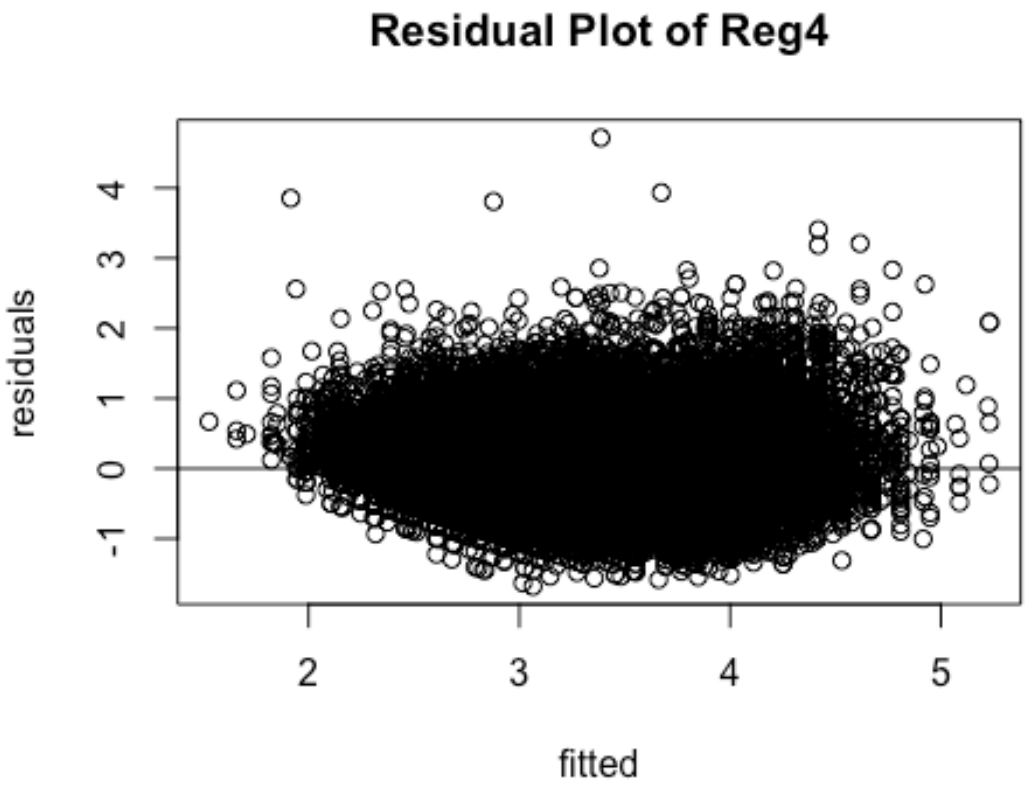
Regression

Linear Model with Mixed Effects

```
## [1] 0.3833741

## [1] 0.4752312

## [1] 0.496054

## [1] 0.5090025
```

For the linear models, I strated my regression with the most simple form by just adding one variable to the right-side of the regression, and developed the model by adding new variables and interaction term.

By comparing the R-squared value, I concluded that reg4 is the most fitted model within these linear models, althougn the R-squared value is 0.509 which is not considerably high, which indicates that the model explains 51% of the variations of the dataset.



The above is the residual plot of reg4. Most of the points are balanced distributed on the top and bottom side of the line cross zero, although there are a few points which are pretty away from zero. Personally this residual plot looks pretty good to me, and indicates that the model reg4 is pretty good fit.

Multilevel regression

Fit a varying intercept model with lmer

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(price) ~ (1 | variety) + (1 | country)
##    Data: wine2
##
## REML criterion at convergence: 155430.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.6472 -0.6657 -0.0386  0.5746  7.8554
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  country  (Intercept) 0.11939  0.3455
##  variety  (Intercept) 0.07077  0.2660
```

```
##  Residual             0.33987  0.5830
## Number of obs: 88223, groups:  country, 41; variety, 20
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  3.14285    0.08319   37.78
```

The above model was created by using the fixed effect "points" to predict price, controlling for by-variety and by-country variability.

From the random effects output, the sd column measures of how much variability in the dependent measure that is due to the random effects "Variety" and "Country". "Residual" which stands for the variability that's not due to the random effects.

From the fixed effects output, the coefficient for points is 5.4, which indicates that a positive 5 times change in price if the points increase by one unit.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(price) ~ points + (1 | variety) + (1 | country)
##    Data: wine2
##
## REML criterion at convergence: 117511.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.5598 -0.6779 -0.0697  0.5873  9.9784
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  country  (Intercept) 0.06793  0.2606
##  variety  (Intercept) 0.04215  0.2053
##  Residual             0.22112  0.4702
## Number of obs: 88223, groups:  country, 41; variety, 20
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept) -7.1435041  0.0793086  -90.07
## points       0.1170682  0.0005378  217.70
##
## Correlation of Fixed Effects:
##        (Intr)
## points -0.596
```

I developed the model by adding a fixed effect term "points". Note that compared to our earlier model without the fixed effect "points", the variation that's associated with the random effect "Variety" and "Country" dropped considerably. This is because the variation that's due to points was confounded with the variation that's due to variety and country. The model didn't know about points, creating relatively larger residuals. Now that we have added the fixed effect of points, we have shifted a

large amount of the variance that was previously in the random effects component to the fixed effects component

From the fixed effects output, the coefficient for points is 0.117, which indicates that a positive 11.7% change in price if the points increase by one unit.

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: log(price) ~ points + (1 | variety) + (1 | country/province)
##    Data: wine2
##
## REML criterion at convergence: 110743.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9205 -0.6681 -0.0633  0.5911 10.7220
##
## Random effects:
##  Groups           Name        Variance Std.Dev.
##  province:country (Intercept) 0.04922  0.2218
##  country          (Intercept) 0.03855  0.1963
##  variety          (Intercept) 0.04155  0.2038
##  Residual                     0.20358  0.4512
## Number of obs: 88223, groups:
## province:country, 359; country, 41; variety, 20
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) -6.6786309  0.0758325  -88.07
## points       0.1119457  0.0005294  211.45
##
## Correlation of Fixed Effects:
##        (Intr)
## points -0.612
```

For the above regression, I fitted the nested group effect terms. Here the (1|country/province) says that we want to fit a mixed effect term for varying intercepts 1| by country, and for province that are nested within country.

For the random effects output, there is still a shift in variance that was in country and variety to a new added random effect term.

From the fixed effects output, there is still a 11% positive change in price if the points increase by one unit.

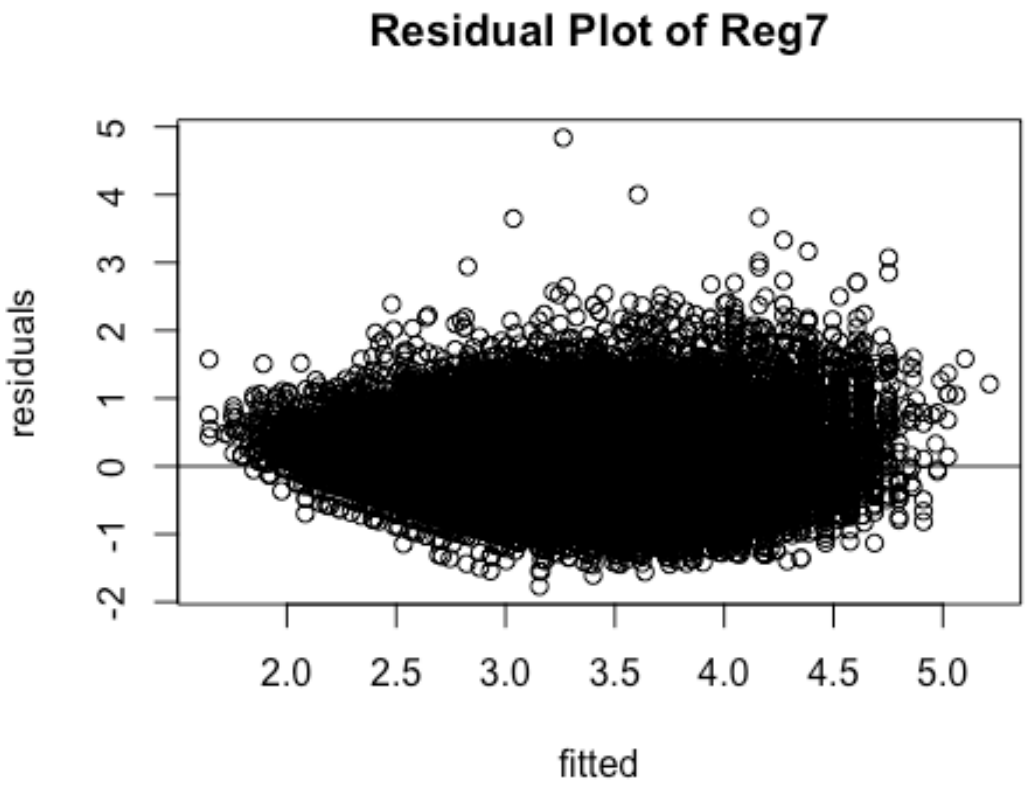
```
##      df      AIC
## reg5  4 155438.2
## reg6  5 117521.8
## reg7  6 110755.4
```

To determine which is the best model, I compared the AIC, and concluded that reg7 is most fitted as it has the lowest AIC value.

To check whether the model did develop or not, I conducted hypothesis test as followings:

```
## Data: wine2
## Models:
## reg.intercept.null: log(price) ~ (1 | variety) + (1 | country)
## reg.intercept.model: log(price) ~ points + (1 | variety) + (1 | coun
try/province)
##                     Df    AIC    BIC logLik deviance Chisq Chi Df
## reg.intercept.null   4 155435 155473 -77714   155427
## reg.intercept.model  6 110738 110795 -55363   110726 44701      2
##                     Pr(>Chisq)
## reg.intercept.null
## reg.intercept.model  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0
```

The results from likelihood ratio test indicated significance, as p-value is very small, with sufficient decimals, could be extremely close to 0. Conclude that the null hypothesis is rejected, that the likelihood of two models are not equivalent. reg7 appeared to be a better model.



From the residual plot above, which indicates the model is good, cause the points are balanced distributed around zero. Which matched the output from the above regression, the residual which stands for the variability that's not due to the random effects is pretty low.

Fit a varying slope model with lmer

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: log(price) ~ points + (1 + points | country/province)
##    Data: wine2
##
##      AIC      BIC   logLik deviance df.resid
## 117334.0 117418.5 -58658.0 117316.0    88214
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9240 -0.6757 -0.0676  0.5971 10.4311
##
## Random effects:
##  Groups           Name        Variance  Std.Dev. Corr
##  province:country (Intercept) 4.2541324 2.06255
##                   points      0.0005641 0.02375  -0.99
##  country          (Intercept) 1.2814751 1.13202
##                   points      0.0001545 0.01243  -0.99
##  Residual                     0.2192037 0.46819
## Number of obs: 88223, groups:  province:country, 359; country, 41
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) -6.967942   0.350128   -19.9
## points       0.114772   0.003958    29.0
##
## Correlation of Fixed Effects:
##        (Intr)
## points -0.994
## convergence code: 1
## Model failed to converge with max|grad| = 23.5059 (tol = 0.002, comp
onent 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
## Model is nearly unidentifiable: large eigenvalue ratio
##  - Rescale variables?
```

The notation "(1+points|country/province)" means that the model is expected to differ baseline-levels of price (the intercept, represented by 1) as well as differ country, and for province that are nested within country.

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula:
## log(price) ~ points + (1 | variety) + (1 + points | country/province)
##    Data: wine2
```

```
##
##      AIC      BIC   logLik deviance df.resid
## 107598.2 107692.0 -53789.1 107578.2    88213
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0570 -0.6561 -0.0513  0.5933 10.9236
##
## Random effects:
##  Groups           Name        Variance  Std.Dev. Corr
##  province:country (Intercept) 2.694e+00 1.641294
##                   points      3.543e-04 0.018823 -0.99
##  country          (Intercept) 6.838e-01 0.826897
##                   points      9.141e-05 0.009561 -0.98
##  variety          (Intercept) 4.390e-02 0.209515
##  Residual                     1.961e-01 0.442798
## Number of obs: 88223, groups:
## province:country, 359; country, 41; variety, 20
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) -6.418499   0.285710  -22.47
## points       0.108835   0.003238   33.61
##
## Correlation of Fixed Effects:
##        (Intr)
## points -0.978
## convergence code: 1
## Model failed to converge with max|grad| = 16.0235 (tol = 0.002, comp
onent 1)
## Model is nearly unidentifiable: very large eigenvalue
##  - Rescale variables?
```

For the random effect output, the variance of all the terms are extremly large, and the residual is huge.

For the fixed effect output, the coefficient of points is 0.109, which indicates a positive change in points would lead to 10.9% increase in the price.

```
##      df      AIC
## reg8  9 117334.0
## reg9 10 107598.2
```

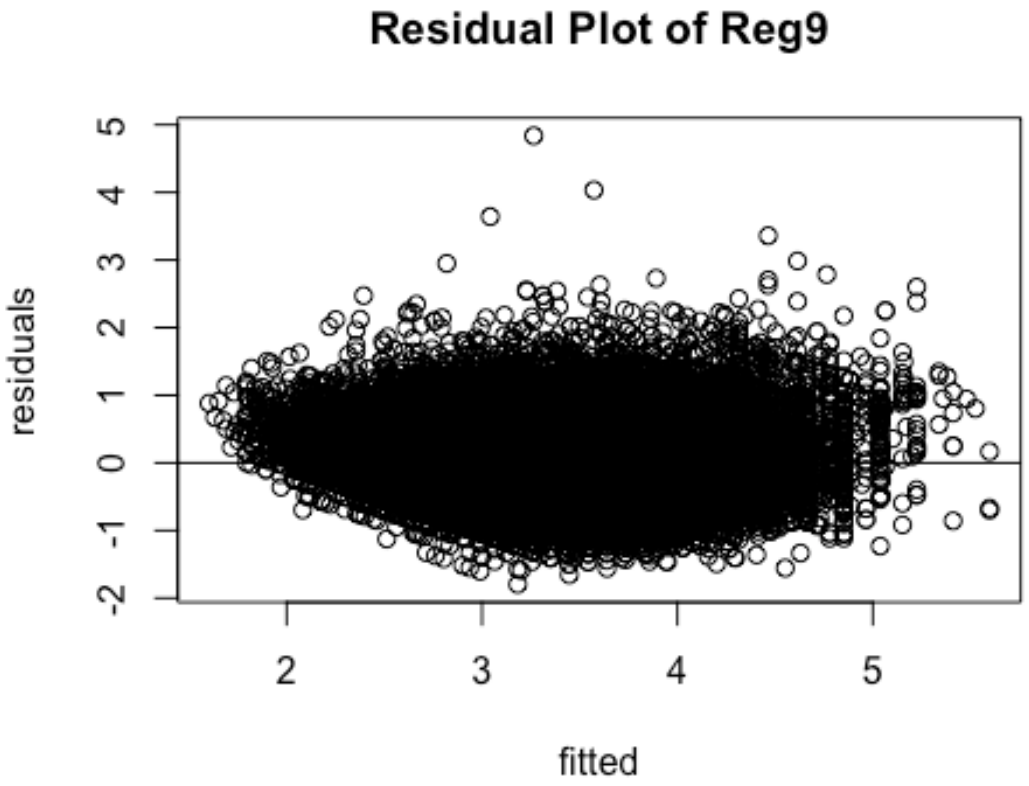
By comparing AIC, concluded that reg9 is most fitted as it has the lowest AIC value.

To check whether the model did develop or not, I conducted hypothesis test as followings:

```
## Data: wine2
## Models:
## reg8: log(price) ~ points + (1 + points | country/province)
```

```
## reg9: log(price) ~ points + (1 | variety) + (1 + points | country/pr
ovince)
##      Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## reg8  9 117334 117418 -58658   117316
## reg9 10 107598 107692 -53789   107578 9737.8      1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] 0
```

As p-value is extremely close to 0 in this case. Conclude that the null hypothesis is rejected, that the likelihood of two models are not equivalent. reg9 appeared to be a better model.



From the residual plot above, which again indicate the model is good, cause the points are balanced distributed around zero. Which matched the output from the above regression, the residual which stands for the variability that's not due to the random effects from the output is low.

Conclusion

Before I have made this final report, I tried several regressions, but I found the data has some kind of right-skewness which made the residual plots have many outliers shown, and led me to a wrong conclusion. Therefore I decided to take log for the variable price, to make it better fit with the models. But there is still some limitations of the data, that More relevant variables are required when exploring the regression on the price of wines.

For my future exploration of this dataset, I would divide the dataset into training and testing, to better test whether the regressions and models I have chosen work with the dataset or not.