Midterm Project Yelp Data Challenge

Xiang Zhao

12/2/2017

1. Introduction

Yelp is a website and also an app collecting the informations of business holders and users. With the increasing population from asia, the number of asian restaurants rising quickly and have been more popular than before. For asian restaurants' holders, in order to offer a better quality for customers and making more profits, analyzing the relationship between the rate of restaurants and restaurants' and users' informations is important. So, I collected the data from Yelp using SQL in R and filter four styles of restaurants, Chinese, Japanese, Korean and Southeast Asian restaurants with their attributes like price range and noise level and so on to fit several models to find the relationship between the rate and restaurants' and users' informations.

At the beginning, I collected the data with all the informations and attributes of restaurants and informations of users. I filter the Chinses, Japanese, Korean and Southeast Asian restaurants with all their informations first. Then I filter six attributes:whether a restaurant has free/paid/no WiFi, the price range of the restaurant, the choices of parking of a restaurant, the noise level of a restaurant, whether a restaurant has a TV or not and whether a restaurant has outdoor seating or not. Next I combined the 'attributes' dataset with 'business' dataset, which gives a whole dataset containing all the informations and attributes of four styles of restaurants. Later, I filter the stars rated by users from 'review' dataset and join it with the 'user' dataset containing users' informations. Finally I join the two datasets of users& reviews and restaurants' informations into one dataset and named it "yelp" which has 472741 rows and 22 variables after cleaning all the NAs.

After getting the clean dataset, I first did EDA especially on the relationships between predictors and response graphically. Then I designed several models to analyze and check these potential relationships statistically and numerically. Finally I summarized the conclusions.

2. Data & Method

2.1Data source

The data is from yelp: https://www.yelp.com/dataset/challenge.

Due to the huge size of dataset after filtering, I saved the filtered dataset into RDS file and reread it in another Rmd file then analyzing. If you need to see the code of reading and cleaning the data, look at 'Data collecting & cleaning.Rmd'.

2.2Method

Tools: SQL, R, csv, RDS, online searching

Packages & Functions: ggplot2::ggplot, lme4::lmer&glmer, VGAM::vglm

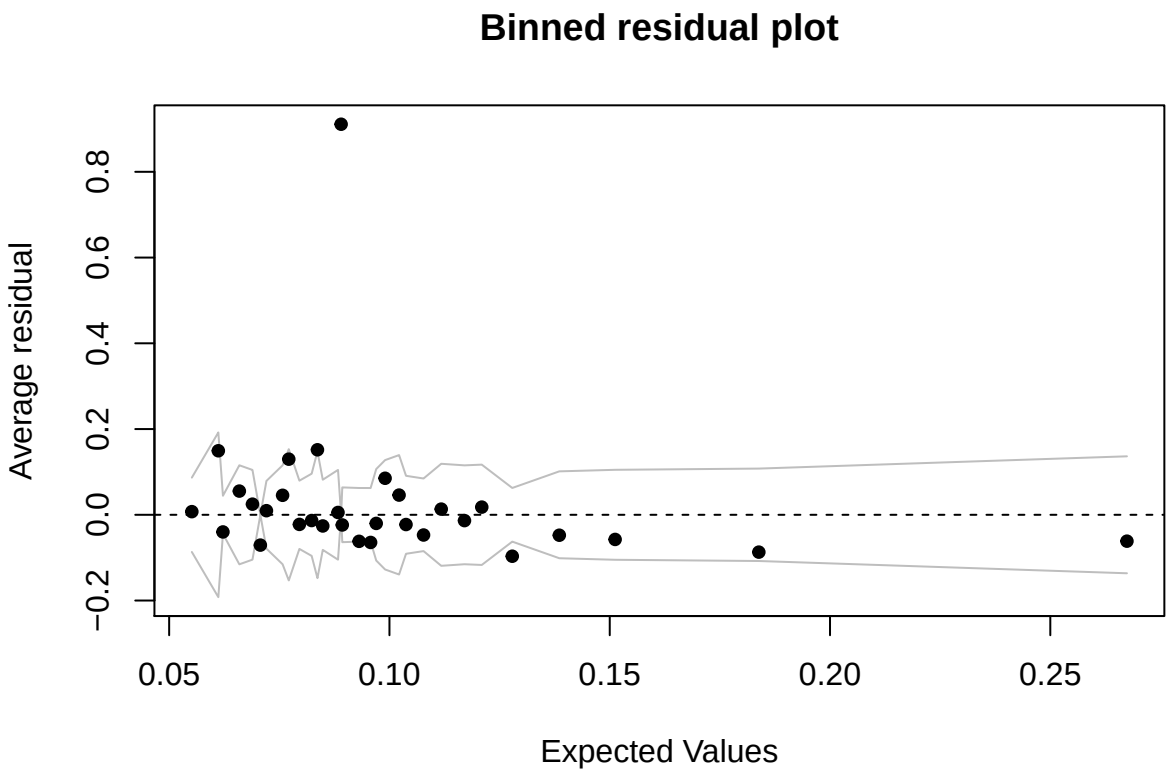
Models: Multilevel linear model, ordered categorical regression

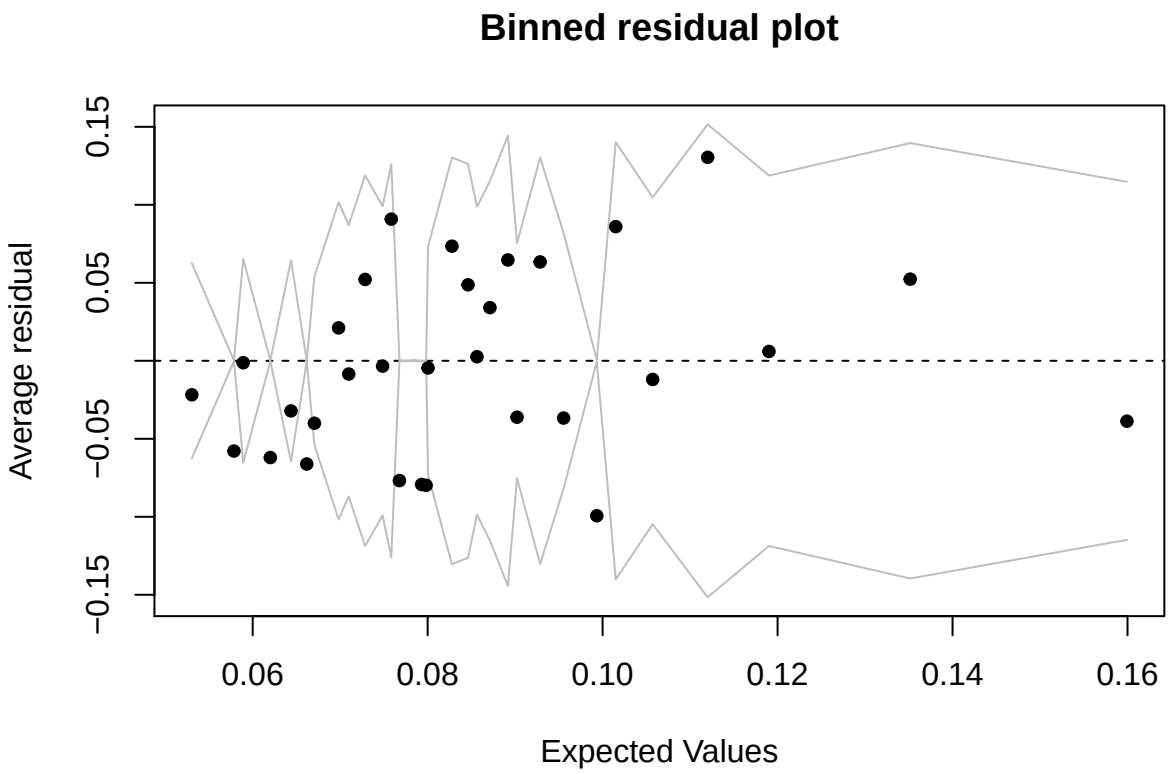
3.EDA

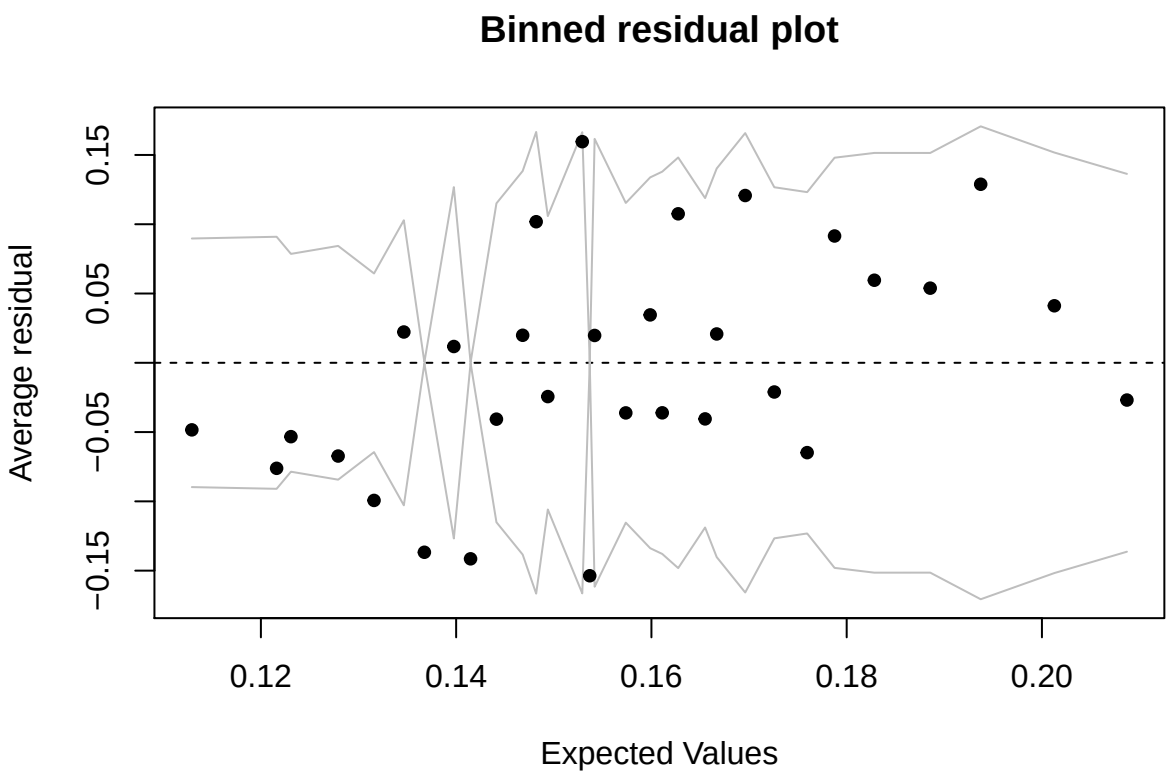
```
##
```

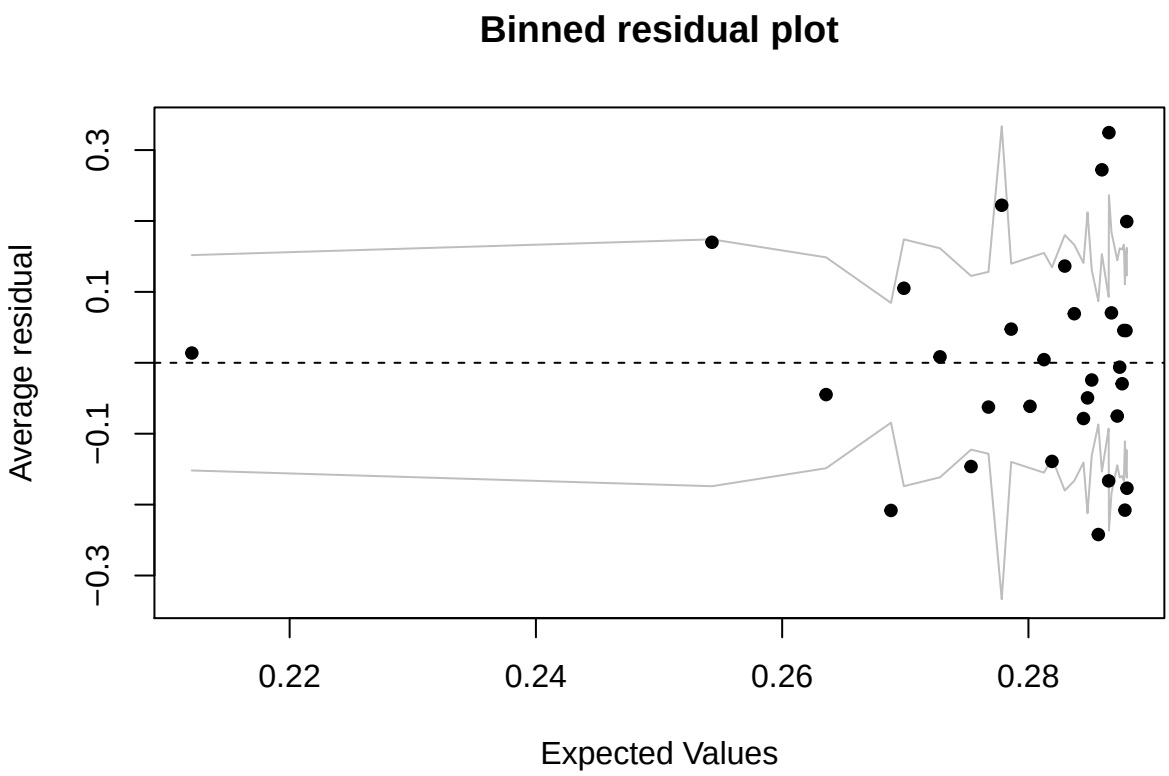
```
## Call:
## vglm(formula = ordered(user_stars) ~ restaurant_style + restaurant_WiFi +
##     restaurant_price_range + garage_parking + street_parking +
##     validated_parking + lot_parking + valet_parking + restaurant_noise_level +
##     restaurant_TV + restaurant_outdoor_seating, family = cumulative(parallel = T),
##     data = mysample)
##
##
## Pearson residuals:
##                     Min      1Q  Median      3Q   Max
## logit(P[Y<=1]) -1.627 -0.2581 -0.1609 -0.1331 3.840
## logit(P[Y<=2]) -1.519 -0.3083 -0.2080 -0.1695 3.759
## logit(P[Y<=3]) -1.846 -0.8893 -0.3182  0.4977 2.532
## logit(P[Y<=4]) -2.357 -1.0789  0.2835  1.0393 1.469
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                   -1.90613    0.26375  -7.227 4.94e-13 ***
## (Intercept):2                   -1.17435    0.25552  -4.596 4.31e-06 ***
## (Intercept):3                   -0.32271    0.25207  -1.280 0.200447
## (Intercept):4                    0.86285    0.25338   3.405 0.000661 ***
## restaurant_styleJapanese        -0.40023    0.15911  -2.515 0.011891 *
## restaurant_styleKorean          -0.28608    0.23480  -1.218 0.223066
## restaurant_styleSoutheast Asian -0.33725    0.15442  -2.184 0.028959 *
## restaurant_WiFino                0.16051    0.13348   1.203 0.229149
## restaurant_WiFipaid              1.15881    1.26271   0.918 0.358766
## restaurant_price_range2          0.07833    0.13925   0.563 0.573770
## restaurant_price_range3         -0.35994    0.39185  -0.919 0.358326
## restaurant_price_range4         -0.35727    0.72859  -0.490 0.623879
## garage_parking true              0.49044    0.25659   1.911 0.055957 .
## street_parking true             -0.15167    0.16952  -0.895 0.370942
## validated_parking true           0.24436    1.27552   0.192 0.848077
## lot_parking true                -0.32096    0.16267  -1.973 0.048487 *
## valet_parking true              -0.31878    0.38965  -0.818 0.413298
## restaurant_noise_levelloud       0.48586    0.28760   1.689 0.091150 .
## restaurant_noise_levelquiet     -0.03553    0.15693  -0.226 0.820878
## restaurant_noise_levelvery_loud  1.34722    0.53806   2.504 0.012286 *
## restaurant_TV1                  -0.17297    0.12379  -1.397 0.162329
## restaurant_outdoor_seating1      0.23386    0.15361   1.522 0.127895
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of linear predictors:  4
##
## Names of linear predictors:
## logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])
##
## Residual deviance: 2874.301 on 3978 degrees of freedom
##
## Log-likelihood: -1437.15 on 3978 degrees of freedom
##
## Number of iterations: 8
##
## No Hauck-Donner effect found in any of the estimates
```

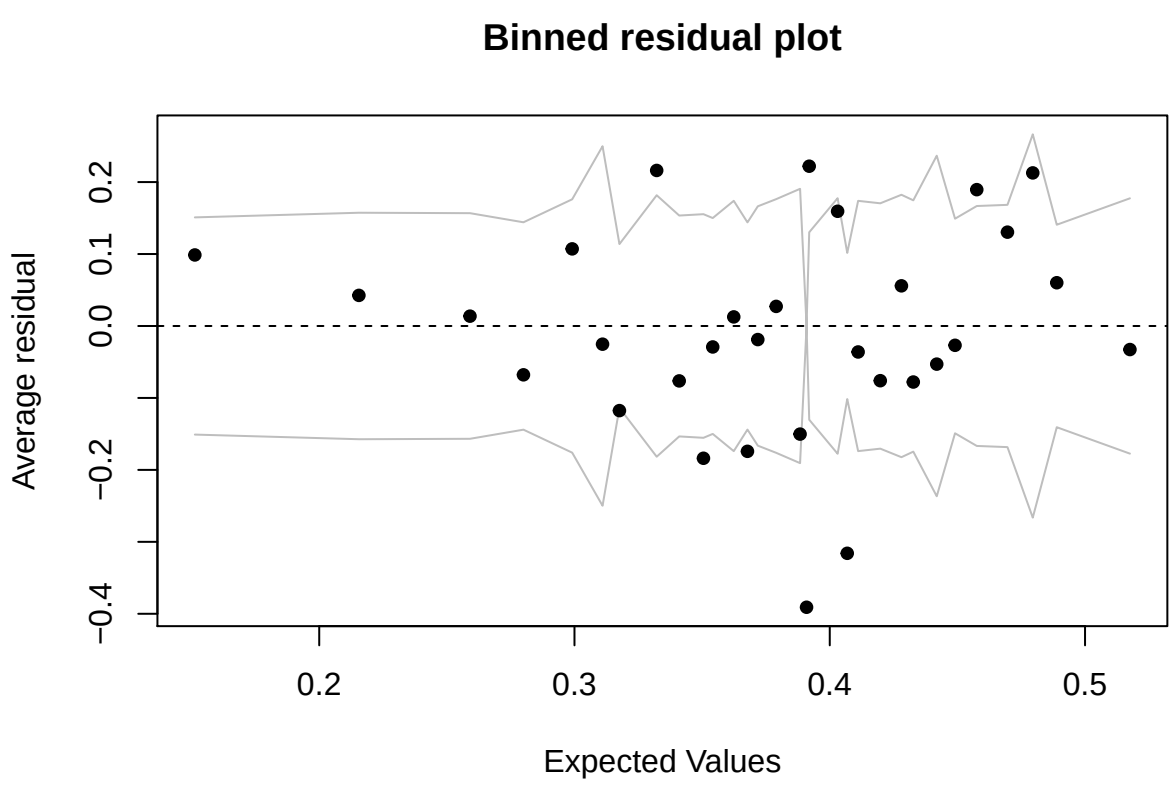
```
##
## Exponentiated coefficients:
##        restaurant_styleJapanese          restaurant_styleKorean
##                       0.6701660                       0.7512032
## restaurant_styleSoutheast Asian               restaurant_WiFino
##                       0.7137285                       1.1741105
##             restaurant_WiFipaid         restaurant_price_range2
##                       3.1861410                       1.0814786
##         restaurant_price_range3         restaurant_price_range4
##                       0.6977217                       0.6995821
##             garage_parking true             street_parking true
##                       1.6330315                       0.8592681
##          validated_parking true                lot_parking true
##                       1.2767988                       0.7254548
##              valet_parking true      restaurant_noise_levelloud
##                       0.7270379                       1.6255658
##     restaurant_noise_levelquiet restaurant_noise_levelvery_loud
##                       0.9650915                       3.8467218
##                  restaurant_TV1     restaurant_outdoor_seating1
##                       0.8411663                       1.2634700
```




Binned residual plot
Average residual
0.15
0.05
−0.05
−0.15
0.06
0.08
0.10
0.12
0.14
0.16
Expected Values

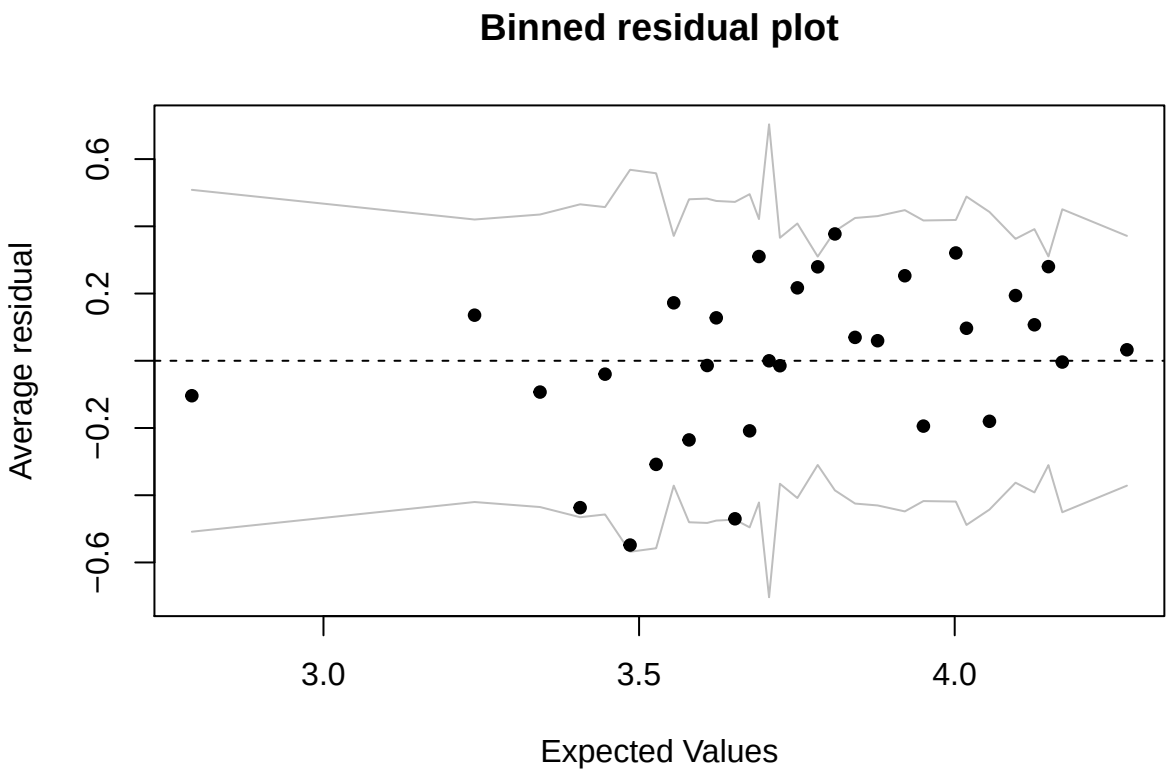

Binned residual plot
Average residual
0.15
0.05
−0.05
−0.15
0.12
0.14
0.16
0.18
0.20
Expected Values


Binned residual plot
Average residual
0.3
0.1
−0.1
−0.3
0.22
0.24
0.26
0.28
Expected Values



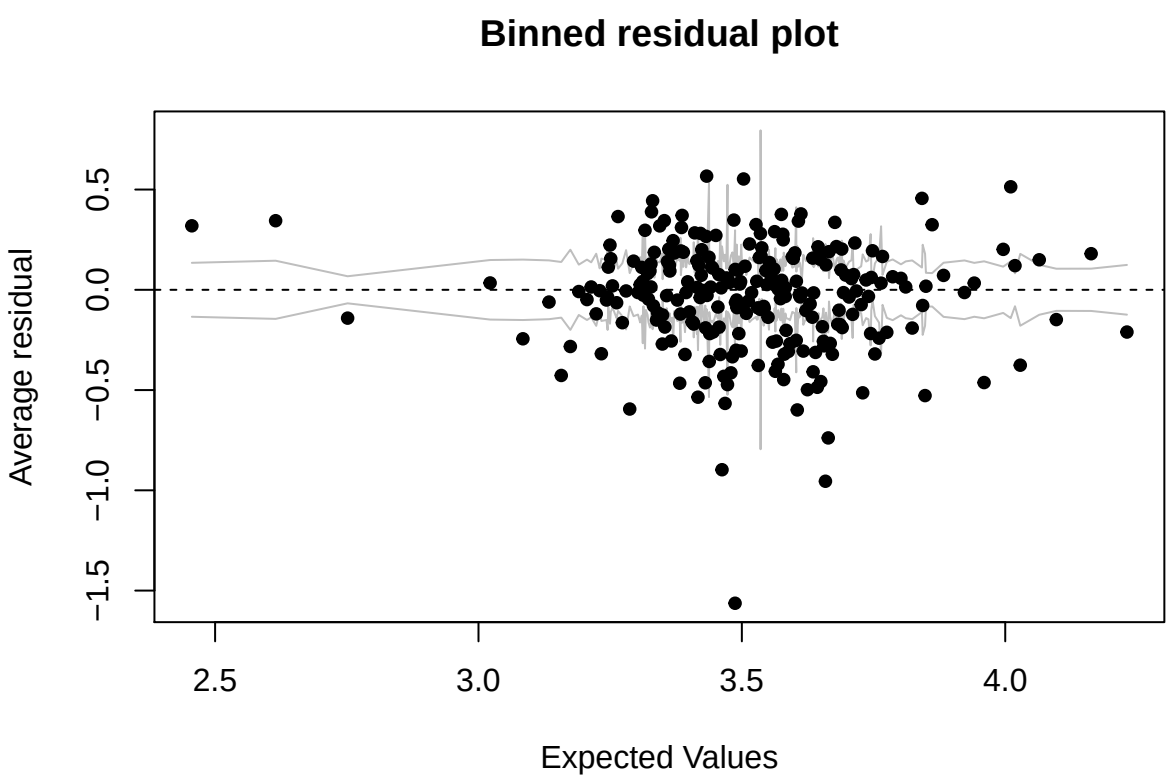

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_style) + (1 | restaurant_city:restaurant_state)
##    Data: mysample
##
## REML criterion at convergence: 3359.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.5719 -0.5575  0.2371  0.7779  1.8089
##
## Random effects:
##  Groups                           Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.067905 0.26059
##  restaurant_style                 (Intercept) 0.004812 0.06937
##  Residual                                     1.635423 1.27884
## Number of obs: 1000, groups:
## restaurant_city:restaurant_state, 71; restaurant_style, 4
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    3.512120   0.184898  18.995
## restaurant_WiFino             -0.014557   0.095237  -0.153
```

```
## restaurant_WiFipaid             -0.982724   0.928693  -1.058
## restaurant_price_range2         -0.071364   0.097485  -0.732
## restaurant_price_range3          0.307248   0.272404   1.128
## restaurant_price_range4          0.223921   0.508274   0.441
## garage_parking true             -0.429972   0.186292  -2.308
## street_parking true              0.220212   0.128603   1.712
## validated_parking true          -0.005204   0.922409  -0.006
## lot_parking true                 0.208600   0.120987   1.724
## valet_parking true               0.024811   0.276032   0.090
## restaurant_noise_levelloud      -0.303924   0.209142  -1.453
## restaurant_noise_levelquiet      0.007036   0.111554   0.063
## restaurant_noise_levelvery_loud -1.044144   0.386705  -2.700
## restaurant_TV1                   0.116097   0.089133   1.303
## restaurant_outdoor_seating1     -0.122595   0.113757  -1.078
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##    vcov(x)     if you need it
```



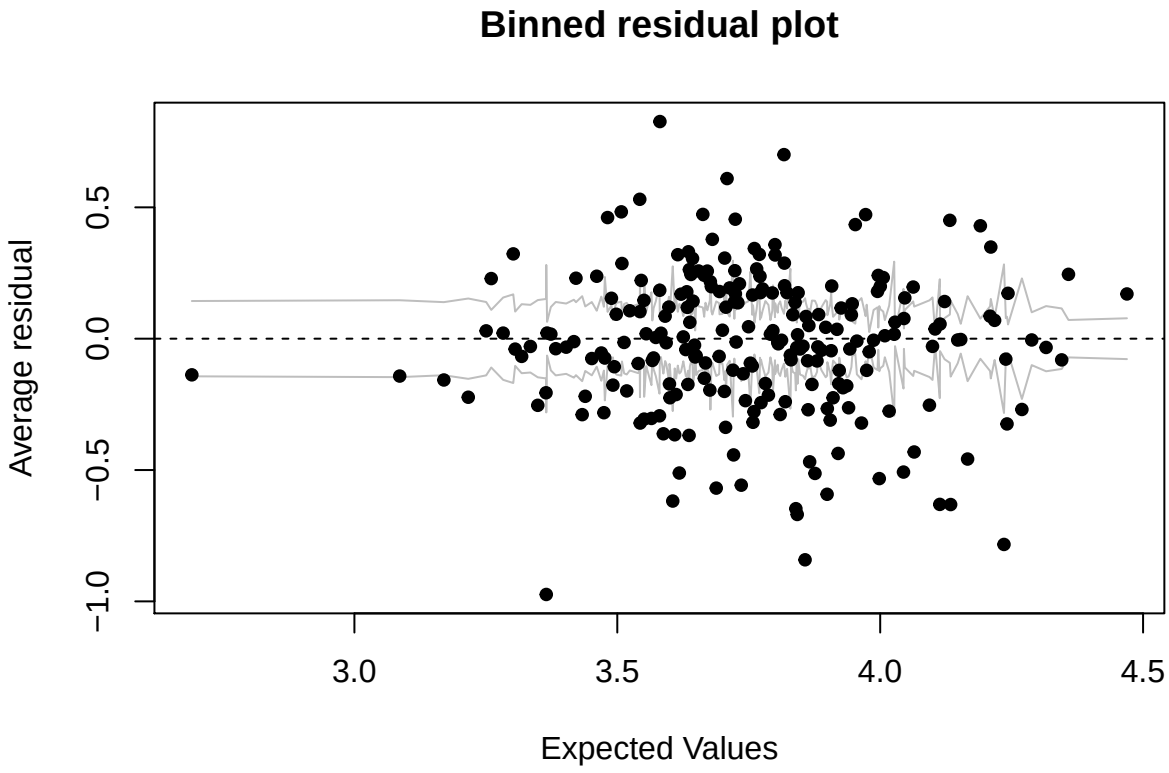

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: chinese
```

```
##
## REML criterion at convergence: 452290.9
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4646 -0.5538  0.3092  0.9359  1.9417
##
## Random effects:
##  Groups                            Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.1033   0.3214
##  Residual                                     1.8244   1.3507
## Number of obs: 131394, groups:  restaurant_city:restaurant_state, 190
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     3.460481   0.031447  110.04
## restaurant_WiFino              -0.074245   0.009681   -7.67
## restaurant_WiFipaid             0.391093   0.124722    3.14
## restaurant_price_range2        -0.076295   0.008556   -8.92
## restaurant_price_range3         0.241182   0.023064   10.46
## restaurant_price_range4         0.292546   0.056898    5.14
## garage_parking true             0.062425   0.016080    3.88
## street_parking true             0.236769   0.014132   16.75
## validated_parking true          0.318027   0.077457    4.11
## lot_parking true                0.169701   0.011287   15.03
## valet_parking true             -0.096321   0.029909   -3.22
## restaurant_noise_levelloud     -0.103661   0.016073   -6.45
## restaurant_noise_levelquiet     0.070449   0.010912    6.46
## restaurant_noise_levelvery_loud -0.838387   0.035536  -23.59
## restaurant_TV1                  0.081026   0.008341    9.71
## restaurant_outdoor_seating1    -0.116363   0.013031   -8.93
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```



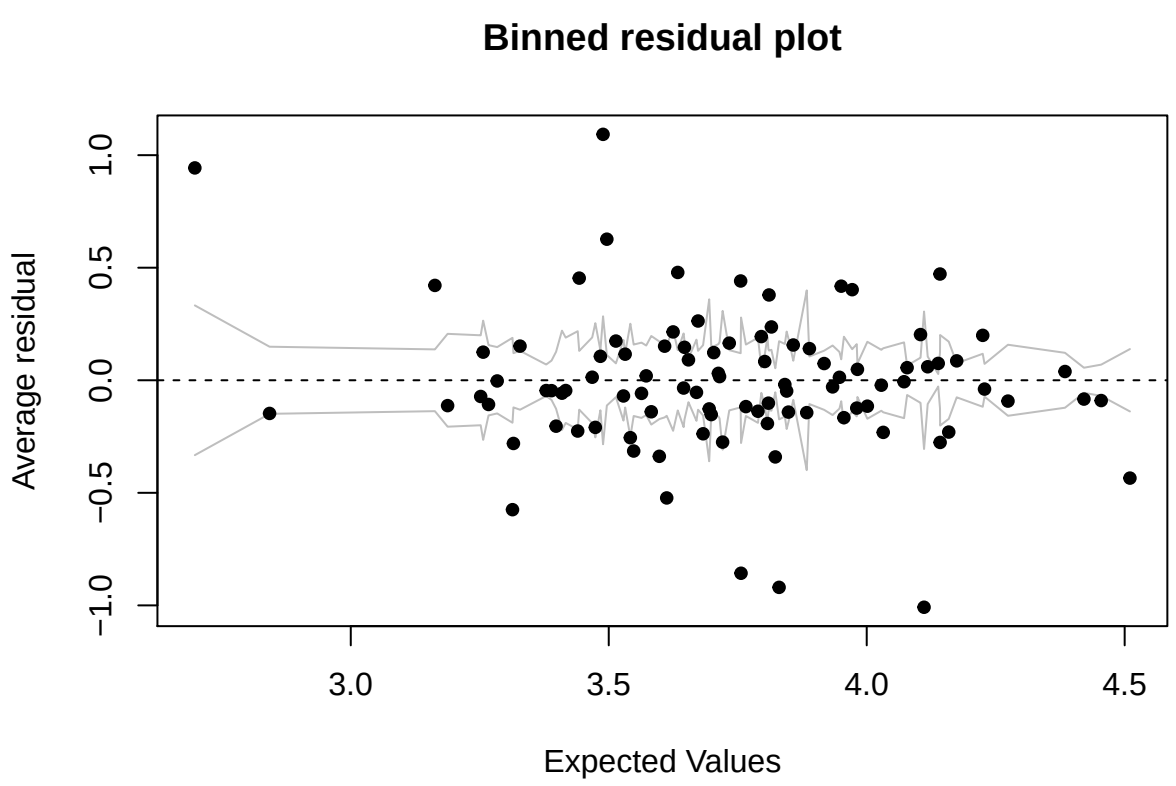

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: japanese
##
## REML criterion at convergence: 498840.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7861 -0.5712  0.2626  0.8048  2.0299
##
## Random effects:
##  Groups                             Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.1465   0.3828
##  Residual                                     1.6089   1.2684
## Number of obs: 150429, groups:  restaurant_city:restaurant_state, 111
##
## Fixed effects:
##                                Estimate Std. Error t value
## (Intercept)                    3.637550   0.041999   86.61
## restaurant_WiFino             -0.125252   0.007699  -16.27
## restaurant_WiFipaid           -0.325194   0.070338   -4.62
## restaurant_price_range2       -0.017203   0.010670   -1.61
```

```
## restaurant_price_range3          0.196647   0.017109   11.49
## restaurant_price_range4          0.124945   0.023910    5.23
## garage_parking true             -0.129792   0.013559   -9.57
## street_parking true              0.136821   0.012295   11.13
## validated_parking true           0.137396   0.042405    3.24
## lot_parking true                 0.032307   0.010401    3.11
## valet_parking true              -0.238354   0.017193  -13.86
## restaurant_noise_levelloud      -0.081484   0.018166   -4.49
## restaurant_noise_levelquiet      0.182440   0.011740   15.54
## restaurant_noise_levelvery_loud  0.169690   0.028072    6.04
## restaurant_TV1                  -0.048817   0.008459   -5.77
## restaurant_outdoor_seating1     -0.057049   0.009343   -6.11
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##    vcov(x)     if you need it
```




```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: korean
##
## REML criterion at convergence: 124142.3
```

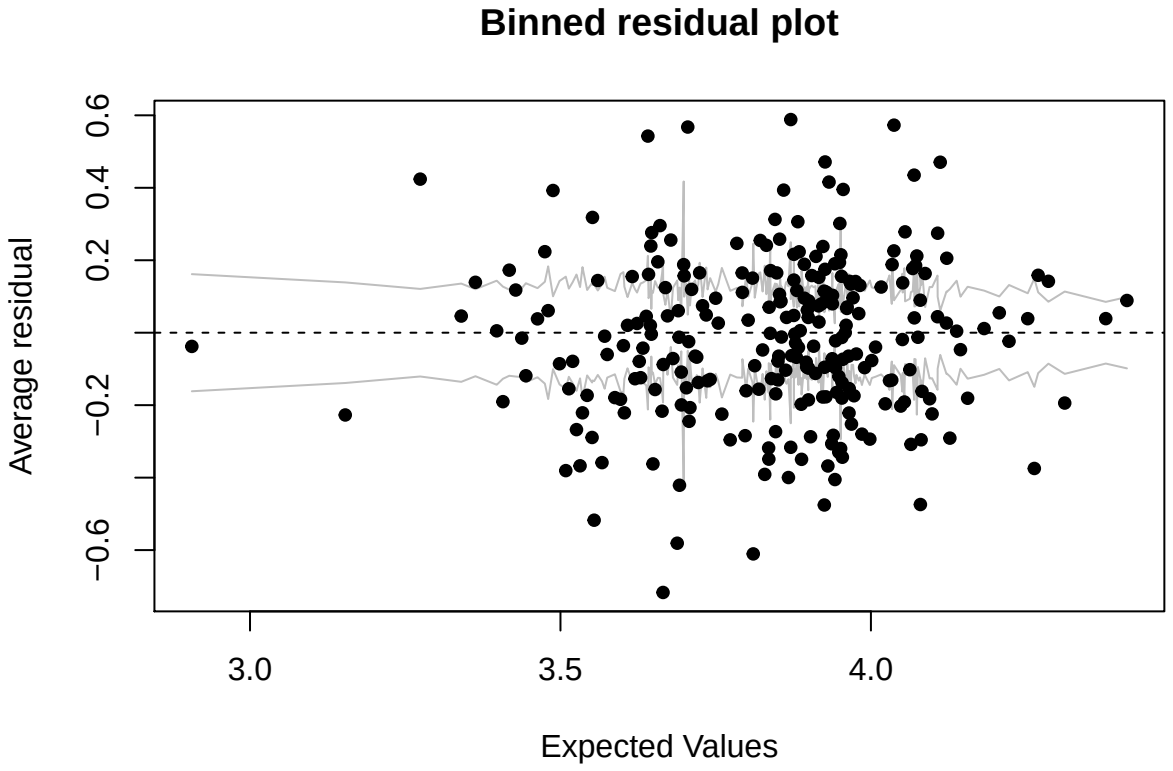
```
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0001 -0.4853  0.2527  0.7220  1.9299
##
## Random effects:
##  Groups                           Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.1203   0.3469
##  Residual                                     1.4232   1.1930
## Number of obs: 38850, groups:  restaurant_city:restaurant_state, 41
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     4.363375   0.065494   66.62
## restaurant_WiFino              -0.315933   0.015895  -19.88
## restaurant_WiFipaid             0.017840   0.086046    0.21
## restaurant_price_range2        -0.281608   0.018035  -15.61
## restaurant_price_range3        -0.723019   0.091844   -7.87
## garage_parking true            -0.690390   0.086335   -8.00
## street_parking true            -0.142824   0.027089   -5.27
## validated_parking true         -0.330166   0.103130   -3.20
## lot_parking true               -0.003439   0.023524   -0.15
## valet_parking true              0.574760   0.075823    7.58
## restaurant_noise_levelloud     -0.134862   0.026442   -5.10
## restaurant_noise_levelquiet     0.022934   0.026623    0.86
## restaurant_noise_levelvery_loud -0.523394   0.168415   -3.11
## restaurant_TV1                 -0.124490   0.016837   -7.39
## restaurant_outdoor_seating1     0.034246   0.023349    1.47
##
## Correlation matrix not shown by default, as p = 15 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```




```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## user_stars ~ restaurant_WiFi + restaurant_price_range + garage_parking +
##     street_parking + validated_parking + lot_parking + valet_parking +
##     restaurant_noise_level + restaurant_TV + restaurant_outdoor_seating +
##     (1 | restaurant_city:restaurant_state)
##    Data: se_asian
##
## REML criterion at convergence: 501276.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7109 -0.5992  0.1649  0.8425  1.8204
##
## Random effects:
##  Groups                             Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.09686  0.3112
##  Residual                                     1.57755  1.2560
## Number of obs: 152068, groups:  restaurant_city:restaurant_state, 140
##
## Fixed effects:
##                                 Estimate Std. Error t value
## (Intercept)                     3.780289   0.032831  115.14
## restaurant_WiFino              -0.000177   0.007656   -0.02
## restaurant_WiFipaid            -0.084038   0.052226   -1.61
## restaurant_price_range2         0.016849   0.007487    2.25
```

```
## restaurant_price_range3          0.278611   0.053409    5.22
## restaurant_price_range4          0.579079   0.563010    1.03
## garage_parking true             -0.280772   0.020810  -13.49
## street_parking true              0.155595   0.011610   13.40
## validated_parking true           0.221032   0.063286    3.49
## lot_parking true                 0.106394   0.010640   10.00
## valet_parking true              -0.205557   0.030688   -6.70
## restaurant_noise_levelloud       0.014001   0.020605    0.68
## restaurant_noise_levelquiet      0.011927   0.008027    1.49
## restaurant_noise_levelvery_loud -0.739897   0.072382  -10.22
## restaurant_TV1                  -0.025635   0.007364   -3.48
## restaurant_outdoor_seating1     -0.003726   0.008817   -0.42

##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```



	Chinese Restaurant	Japanese Restaurant	Southeast Asian Restaurant	Korean Res
Intercept	3.4604814	3.6375503	3.7802886	4.
restaurant_WiFino	-0.0742453	-0.1252521	-0.0001770	-0.
restaurant_WiFipaid	0.3910925	-0.3251938	-0.0840379	0.
restaurant_price_range2	-0.0762948	-0.0172033	0.0168487	-0.
restaurant_price_range3	0.2411818	0.1966469	0.2786113	-0.
restaurant_price_range4	0.2925457	0.1249455	0.5790789	
garage_parking true	0.0624246	-0.1297917	-0.2807717	-0.

	Chinese Restaurant	Japanese Restaurant	Southeast Asian Restaurant	Korean Res
street_parking true	0.2367694	0.1368207	0.1555946	-0.
validated_parking true	0.3180269	0.1373965	0.2210321	-0.
lot_parking true	0.1697010	0.0323071	0.1063941	-0.
valet_parking true	-0.0963215	-0.2383541	-0.2055574	0.
restaurant_noise_levelloud	-0.1036612	-0.0814839	0.0140009	-0.
restaurant_noise_levelquiet	0.0704491	0.1824398	0.0119266	0.
restaurant_noise_levelvery_loud	-0.8383866	0.1696895	-0.7398966	-0.
restaurant_TV1	0.0810256	-0.0488169	-0.0256351	-0.
restaurant_outdoor_seating1	-0.1163634	-0.0570491	-0.0037261	0.

	city:state
Chinese	0.10330
Japanese	0.14650
Korean	0.12030
Southeast Asian	0.09686