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

2.1 Data 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'.

Part of data shows below:

1st-5th columns:

business_id	restaurant_style	restaurant_name	restaurant_city	restaurant_state
iMoFE2g4kDG4FfKLJvk3Jw	Korean	Buk Chang Dong Soon Tofu	North York	ON
iMoFE2g4kDG4FfKLJvk3Jw	Korean	Buk Chang Dong Soon Tofu	North York	ON
iMoFE2g4kDG4FfKLJvk3Jw	Korean	Buk Chang Dong Soon Tofu	North York	ON
iMoFE2g4kDG4FfKLJvk3Jw	Korean	Buk Chang Dong Soon Tofu	North York	ON
iMoFE2g4kDG4FfKLJvk3Jw	Korean	Buk Chang Dong Soon Tofu	North York	ON
iMoFE2g4kDG4FfKLJvk3Jw	Korean	Buk Chang Dong Soon Tofu	North York	ON

6th-9th columns:

restaurant_stars	restaurant_review_count	restaurant_WiFi	restaurant_price_range
4	267	no	1
4	267	no	1
4	267	no	1
4	267	no	1
4	267	no	1
4	267	no	1

10th-13rd columns:

garage_parking	street_parking	validated_parking	lot_parking	valet_parking
false	true	false	false	false
false	true	false	false	false
false	true	false	false	false
false	true	false	false	false
false	true	false	false	false
false	true	false	false	false

15th-17th columns:

restaurant_noise_level	restaurant_TV	restaurant_outdoor_seating
average	0	0
average	0	0

restaurant_noise_level	restaurant_TV	restaurant_outdoor_seating
average	0	0
average	0	0
average	0	0
average	0	0

18th-22nd columns:

user_id	user_stars	user_name	user_review_count	user_average_stars
a4PU5fqFJynStdxMlxRflg	3	Fiona	37	3.55
zxnJHs9eEYfVq76LDTg9JA	5	Fiona	10	3.4
8P2LkzPGV4ID_fE7gJGKGg	3	Nadia	153	3.23
6LjTkzT-hFtf_YWjMMQffQ	4	Naseer	3	2
X2FKoMQOkGr17HOdqo6B0g	2	Christine	27	3.07
edEbC4fEPqq2BhbDaFj6Yw	5	Emily	4	4.25

2.2 Method

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

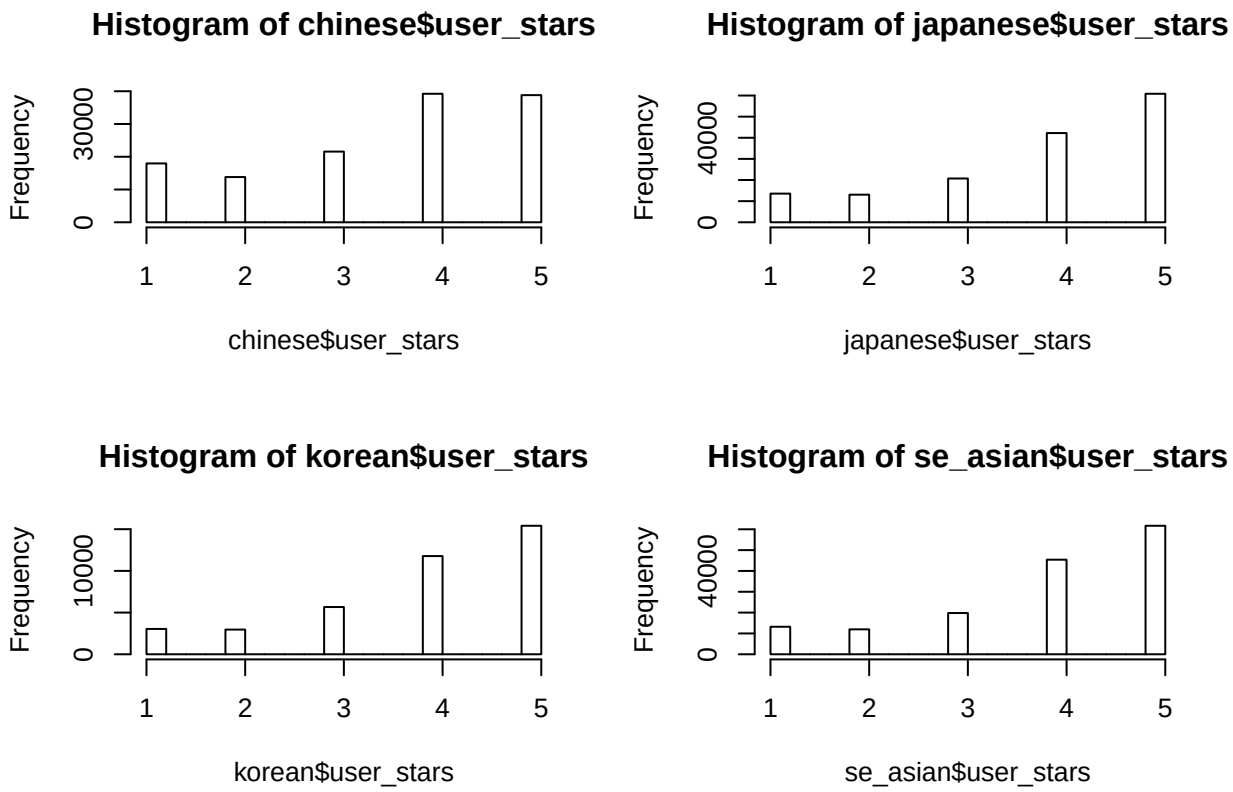
Packages & Functions: ggplot2::ggplot, lme4::lmer&glmer, VGAM::vglm

Models: Multilevel linear model, ordered categorical regression

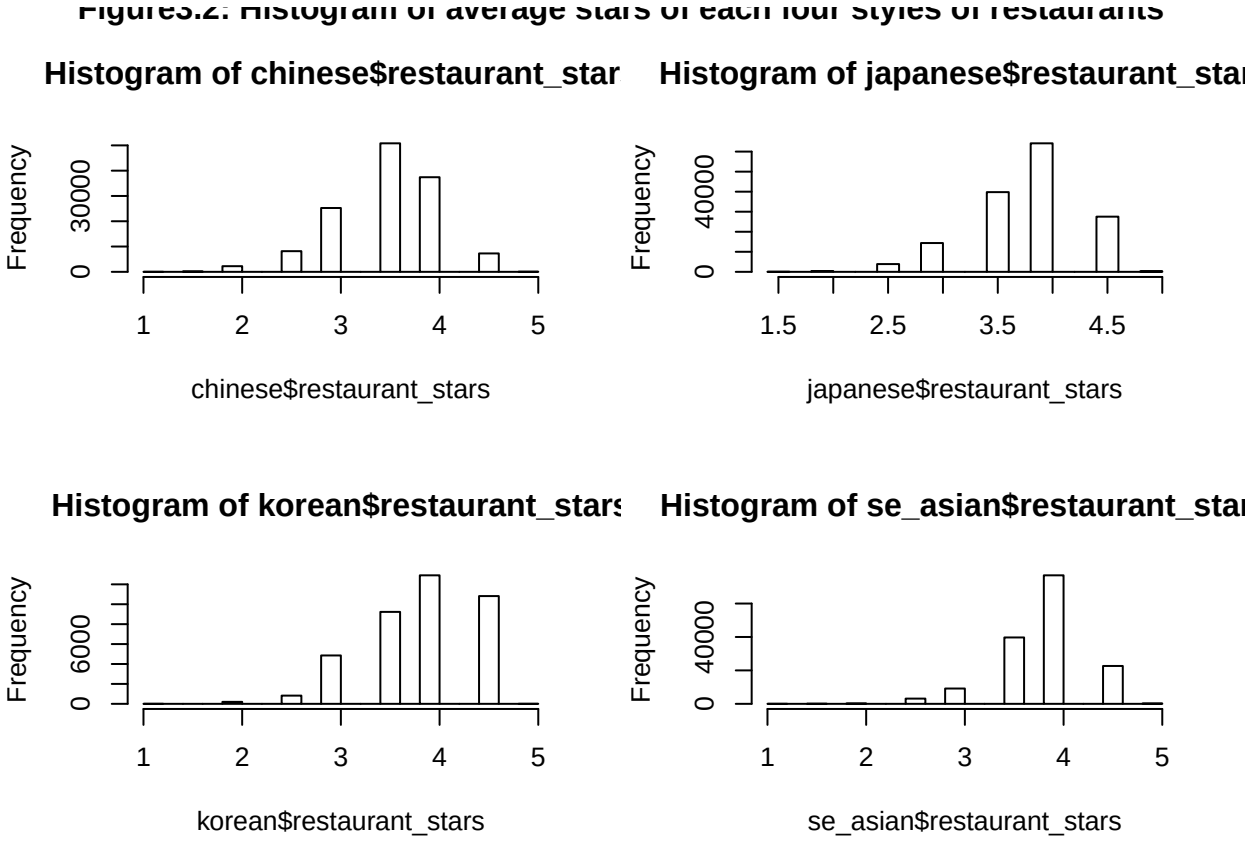
3.EDA

3.1 Histogram of stars rated by users toward each four styles of restaurants

Figure3.1: Histogram of stars rated by users toward each four styles of restaurants



3.2 Histogram of average stars of each four styles of restaurants



3.3 Stacked bar plot of count of stars within each predictor.

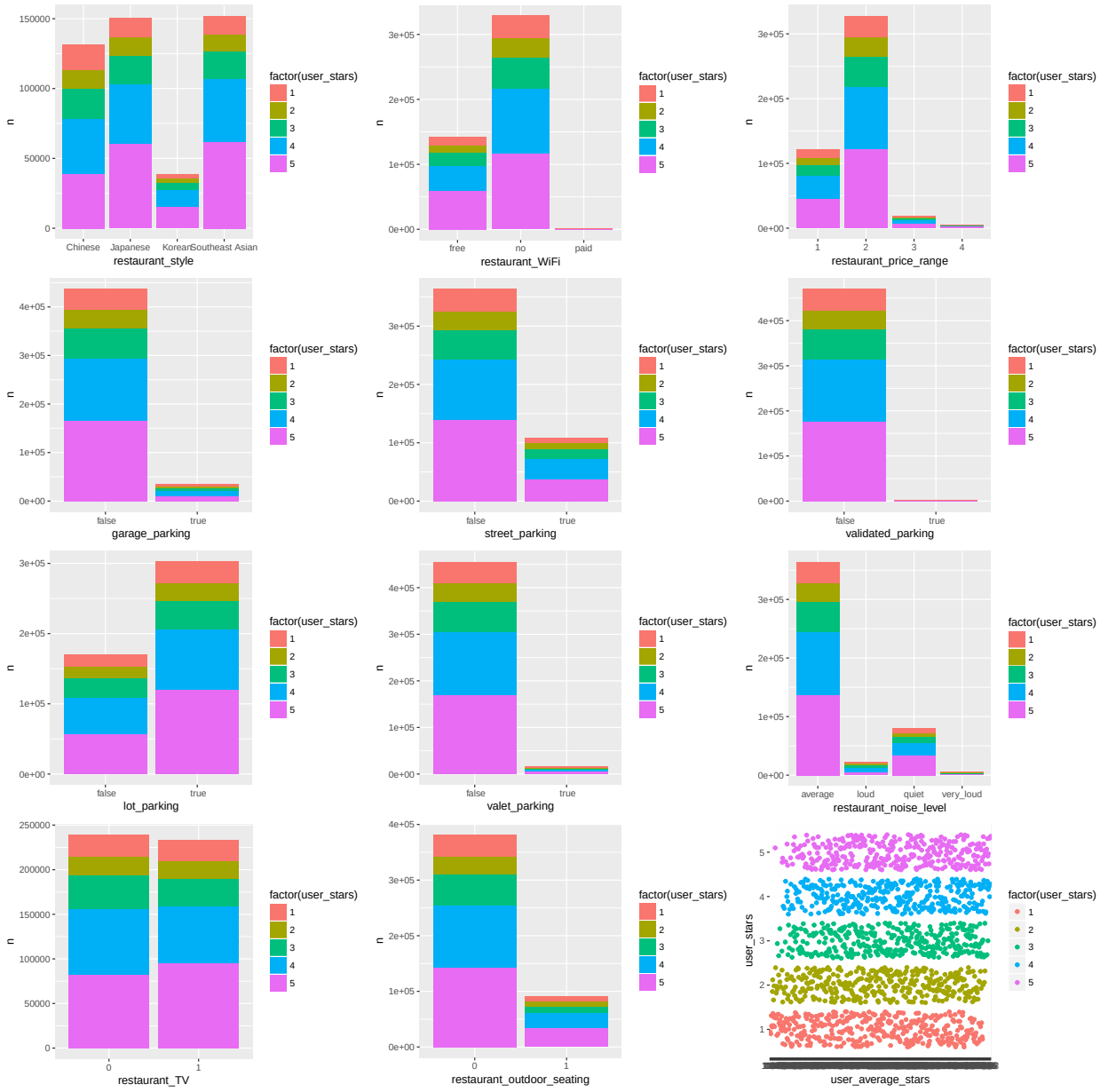


Figure3.3: Stacked bar plot of count of stars within each predictor

3.4 Stacked bar plot of percentage of stars within each predictor.

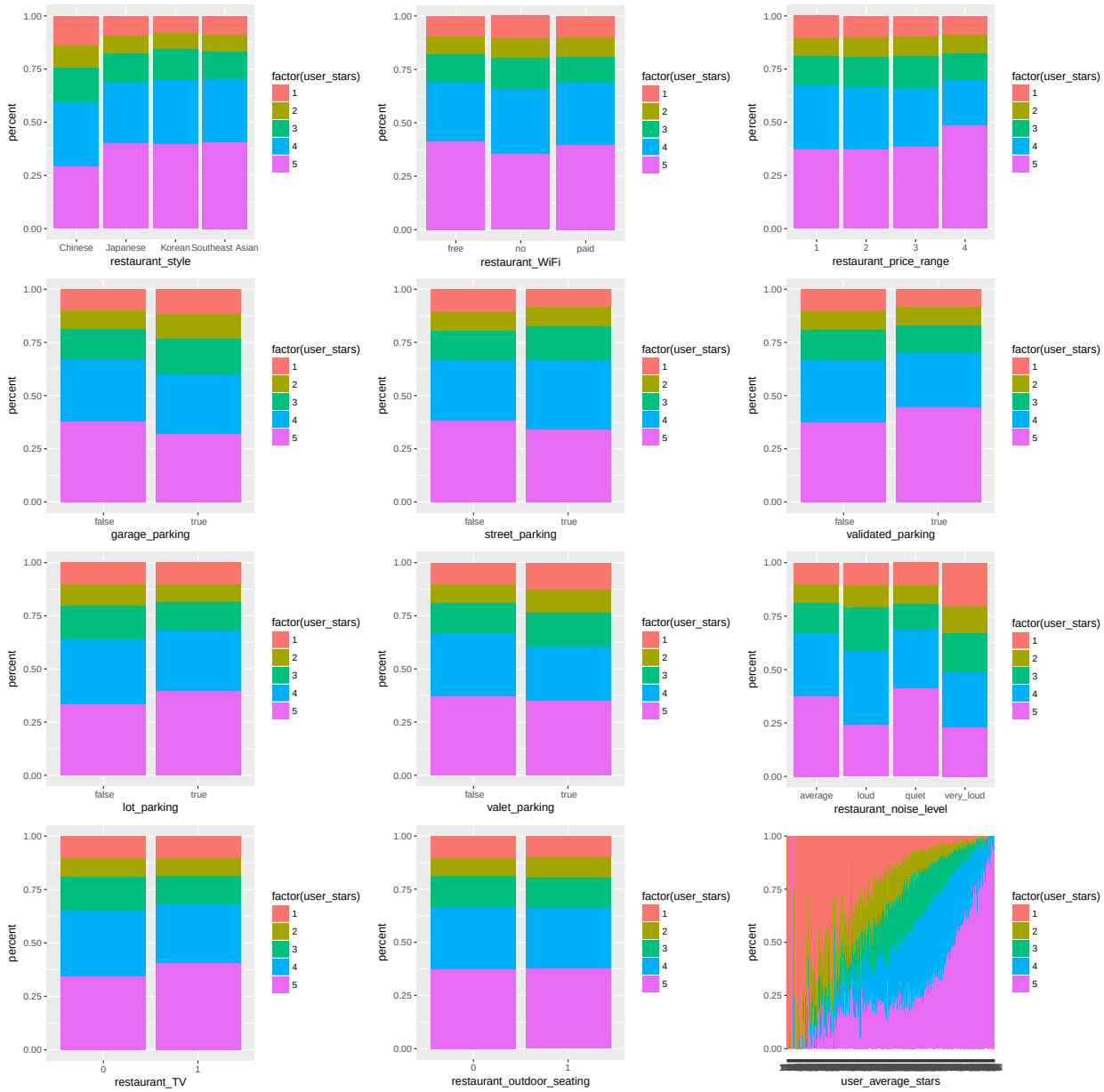


Figure3.4: Stacked bar plot of percentage of stars within each predictor

4.Model fitting & Analysis

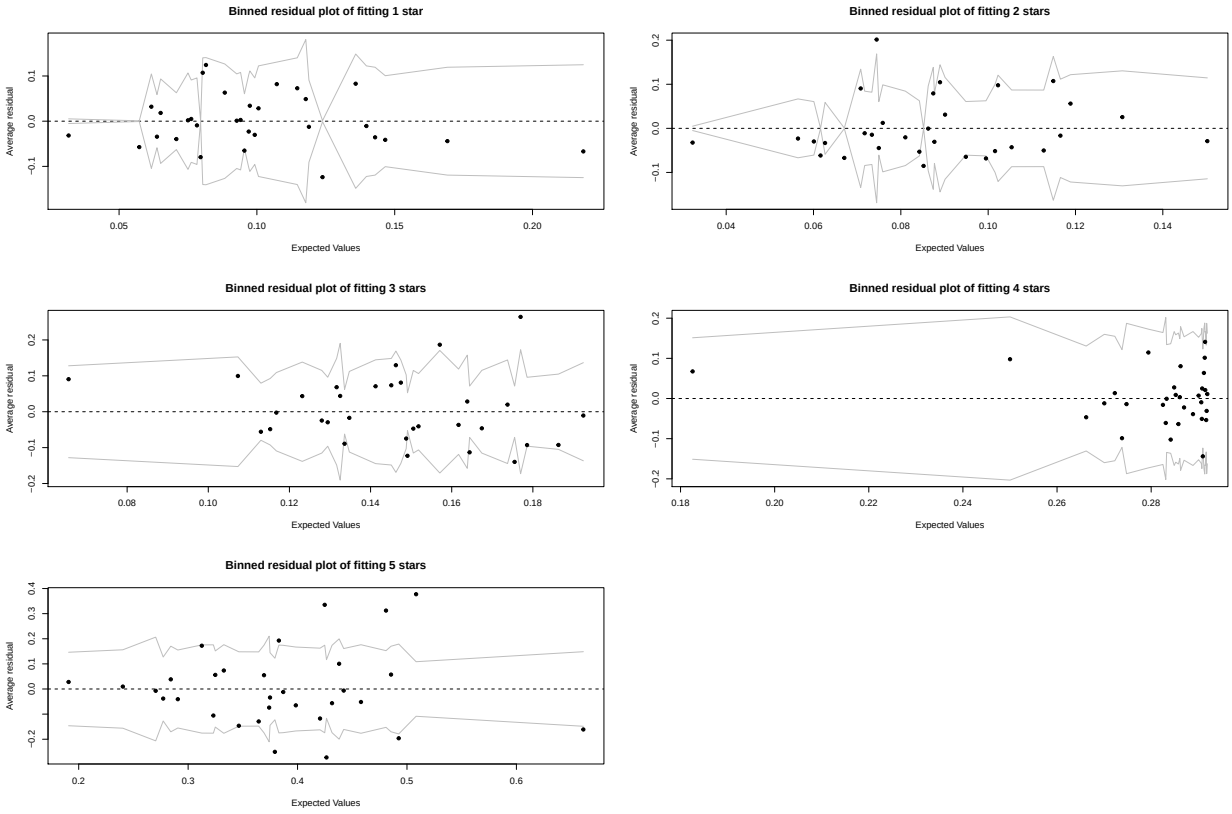
4.1 Categorical Regression

I first fit a categorical regression with the star of a user rating for a certain restaurant as response and the style, price range, noise level, parking conditions and wifi, TV and outdoor seatings of a restaurant as predictors. This model does not have random effects.

```
##
## 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]) -1.597 -0.2348 -0.1637 -0.1326 3.619
## logit(P[Y<=2]) -1.556 -0.3145 -0.2137 -0.1681 3.683
## logit(P[Y<=3]) -1.527 -0.8670 -0.3114  0.4728 4.593
## logit(P[Y<=4]) -2.181 -1.0647  0.2812  1.0207 2.095
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                   -2.10913    0.25063  -8.415  < 2e-16 ***
## (Intercept):2                   -1.36511    0.24161  -5.650  1.6e-08 ***
## (Intercept):3                   -0.57227    0.23765  -2.408  0.01604 *
## (Intercept):4                    0.63077    0.23783   2.652  0.00800 **
## restaurant_styleJapanese        -0.43244    0.16307  -2.652  0.00801 **
## restaurant_styleKorean          -0.59430    0.24698  -2.406  0.01612 *
## restaurant_styleSoutheast Asian -0.45631    0.15079  -3.026  0.00248 **
## restaurant_WiFino                0.23054    0.13292   1.734  0.08283 .
## restaurant_WiFipaid             -1.54477    1.09060  -1.416  0.15665
## restaurant_price_range2          0.03876    0.13680   0.283  0.77693
## restaurant_price_range3         -0.42099    0.40748  -1.033  0.30153
## restaurant_price_range4         -1.69263    0.60037  -2.819  0.00481 **
## garage_parking true              0.24186    0.24285   0.996  0.31929
## street_parking true              0.02020    0.16992   0.119  0.90537
## validated_parking true          -0.62487    0.94590  -0.661  0.50886
## lot_parking true                 0.06953    0.16135   0.431  0.66651
## valet_parking true               0.44600    0.38674   1.153  0.24881
## restaurant_noise_levelloud       0.71740    0.27426   2.616  0.00890 **
## restaurant_noise_levelquiet      0.00129    0.15714   0.008  0.99345
## restaurant_noise_levelvery_loud -0.01185    0.48028  -0.025  0.98031
## restaurant_TV1                  -0.23511    0.12348  -1.904  0.05692 .
## restaurant_outdoor_seating1      0.26808    0.14947   1.793  0.07290 .
## ---
## 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: 2849.735 on 3978 degrees of freedom
##
## Log-likelihood: -1424.867 on 3978 degrees of freedom
##
## Number of iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##        restaurant_styleJapanese          restaurant_styleKorean
##                       0.6489221                       0.5519514
## restaurant_styleSoutheast Asian               restaurant_WiFino
##                       0.6336179                       1.2592794
##             restaurant_WiFipaid         restaurant_price_range2
##                       0.2133602                       1.0395194
##         restaurant_price_range3         restaurant_price_range4
##                       0.6563967                       0.1840354
##             garage_parking true             street_parking true
##                       1.2736114                       1.0204059
##          validated_parking true                lot_parking true
##                       0.5353309                       1.0720052
##              valet_parking true      restaurant_noise_levelloud
##                       1.5620495                       2.0491069
##     restaurant_noise_levelquiet restaurant_noise_levelvery_loud
##                       1.0012911                       0.9882146
##                  restaurant_TV1     restaurant_outdoor_seating1
##                       0.7904866                       1.3074446
```

Figure4.1: Binned residual plots for each category of rated stars

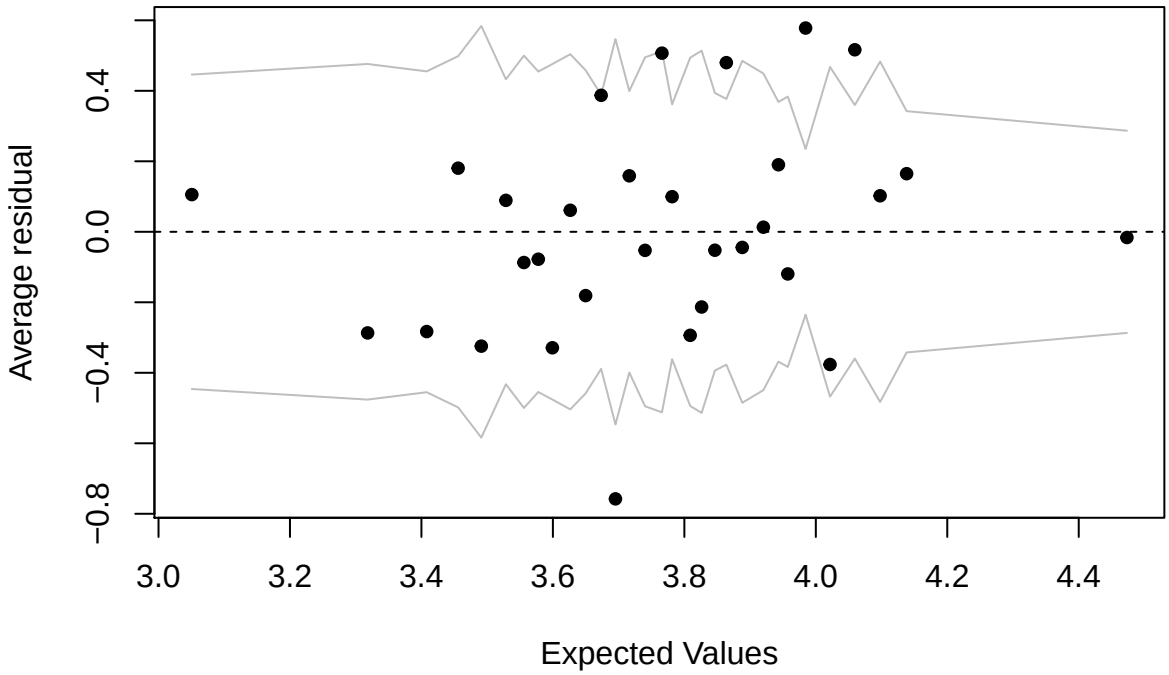


4.2 Multilevel linear model (combine four styles of restaurants and sample the data for regression)

```
## 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: 3374.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4830 -0.6022  0.2289  0.8168  1.5414
##
## Random effects:
##  Groups                          Name        Variance Std.Dev.
##  restaurant_city:restaurant_state (Intercept) 0.03674  0.1917
##  restaurant_style                 (Intercept) 0.02691  0.1640
##  Residual                                     1.66147  1.2890
## Number of obs: 1000, groups:
## restaurant_city:restaurant_state, 74; restaurant_style, 4
##
## Fixed effects:
```

```
##                                 Estimate Std. Error t value
## (Intercept)                      3.81696    0.18796  20.308
## restaurant_WiFino               -0.10928    0.09419  -1.160
## restaurant_WiFipaid              0.67909    0.59101   1.149
## restaurant_price_range2         -0.02667    0.09745  -0.274
## restaurant_price_range3          0.29109    0.28522   1.021
## restaurant_price_range4          0.99215    0.35926   2.762
## garage_parking true             -0.24085    0.17573  -1.371
## street_parking true              0.10205    0.12754   0.800
## validated_parking true           0.45961    0.59723   0.770
## lot_parking true                -0.03836    0.11775  -0.326
## valet_parking true              -0.26709    0.26974  -0.990
## restaurant_noise_levelloud      -0.53001    0.19980  -2.653
## restaurant_noise_levelquiet     -0.04277    0.11223  -0.381
## restaurant_noise_levelvery_loud  0.02472    0.34445   0.072
## restaurant_TV1                   0.12772    0.08822   1.448
## restaurant_outdoor_seating1     -0.18253    0.10940  -1.668
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##   vcov(x)     if you need it
```

Figure4.2: Binned residual plot of fitting rated stars of sample data

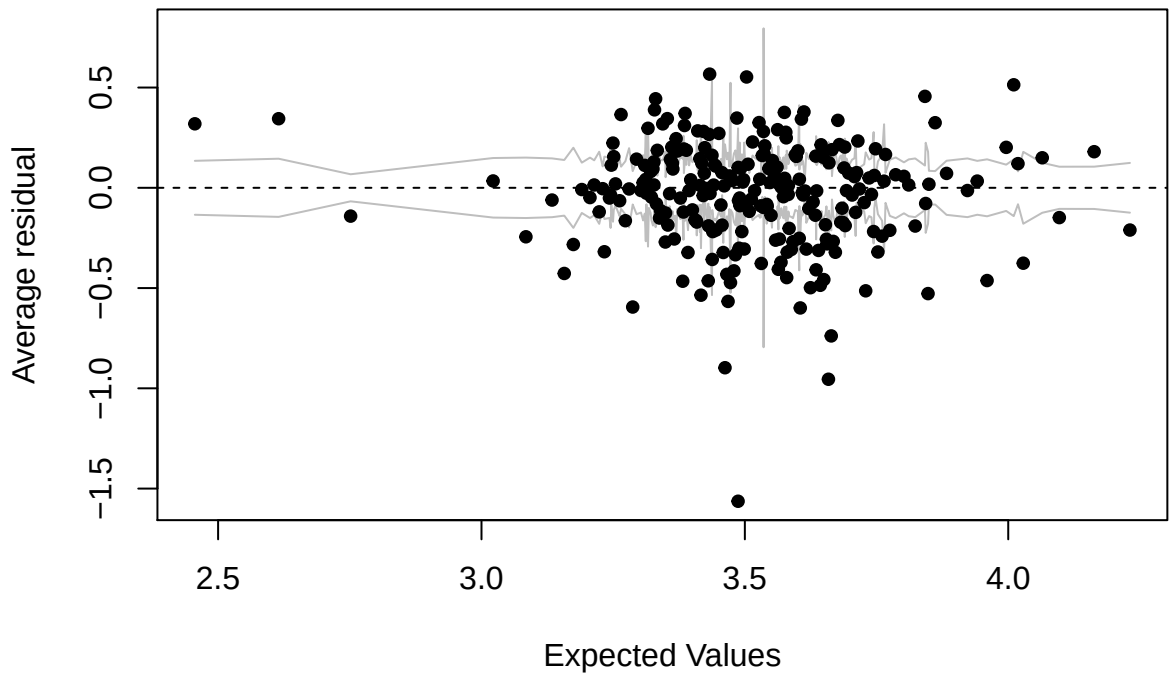


4.3 Multilevel linear model (separate four styles of restaurants and fit model individually)

First I fit the model only with data of Chinese restaurants.

```
## 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
```

Figure4.3: Binned residual plot of fitting rated stars of Chinese restaur

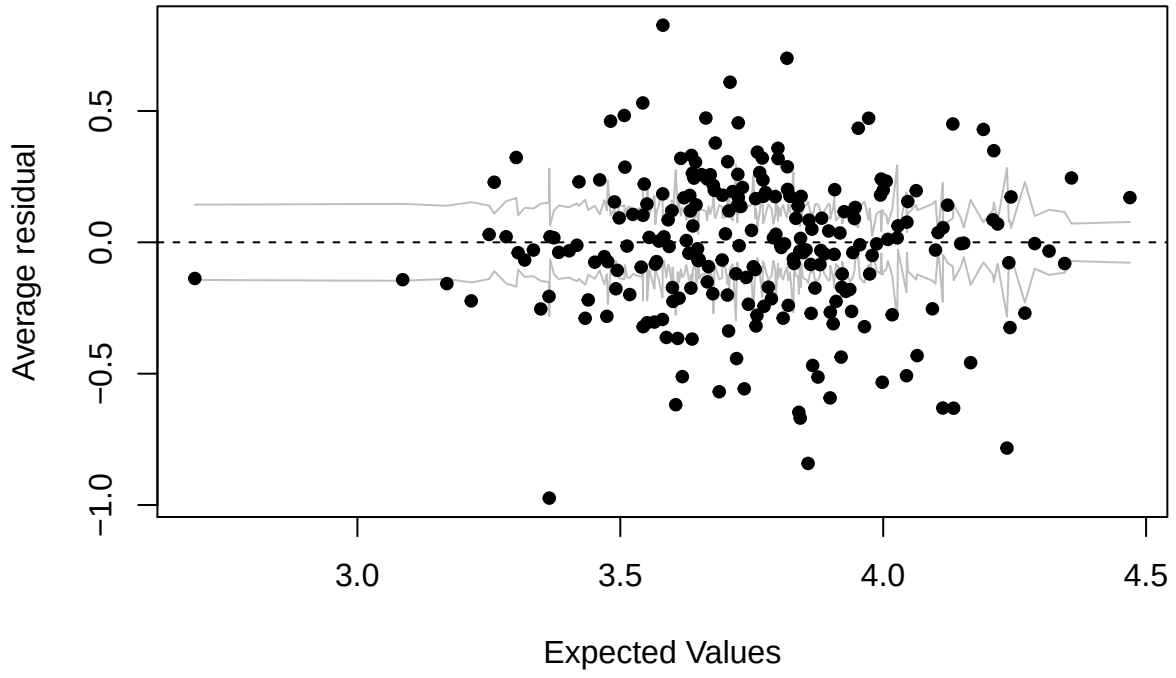


Second I fit the model only with data of Japanese restaurants.

```
## 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
```

Figure4.4: Binned residual plot of fitting rated stars of Japanese restau

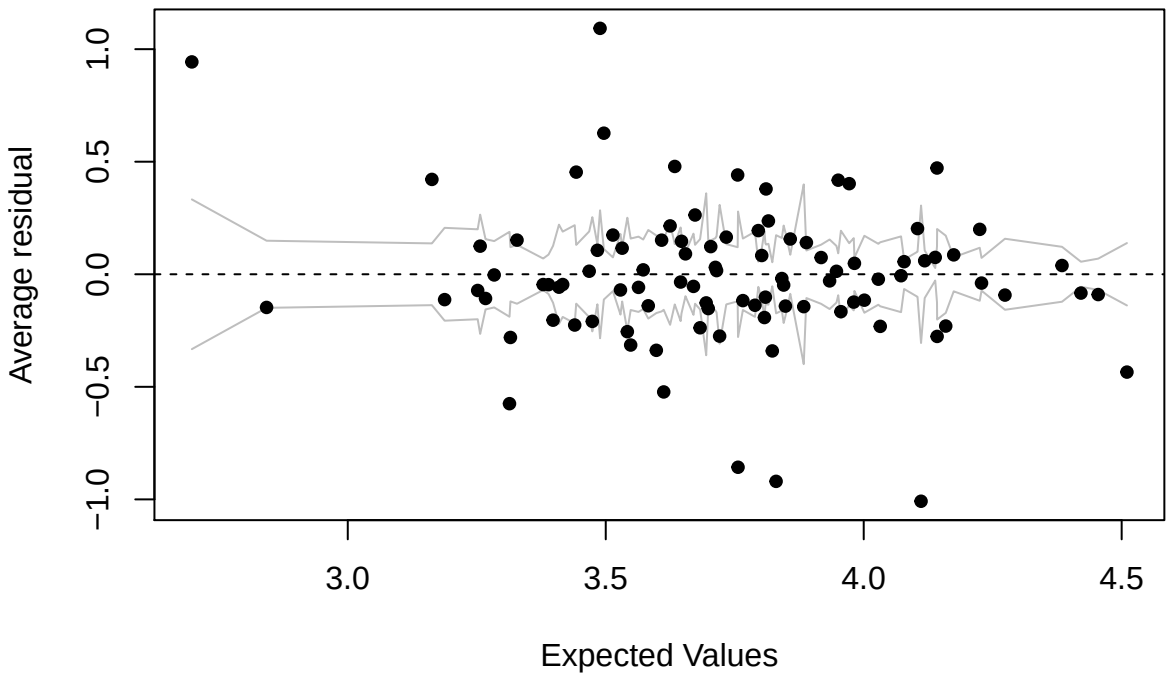


Third I fit the model only with data of Korean restaurants.

```
## 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
```

Figure4.5: Binned residual plot of fitting rated stars of Korean restaura



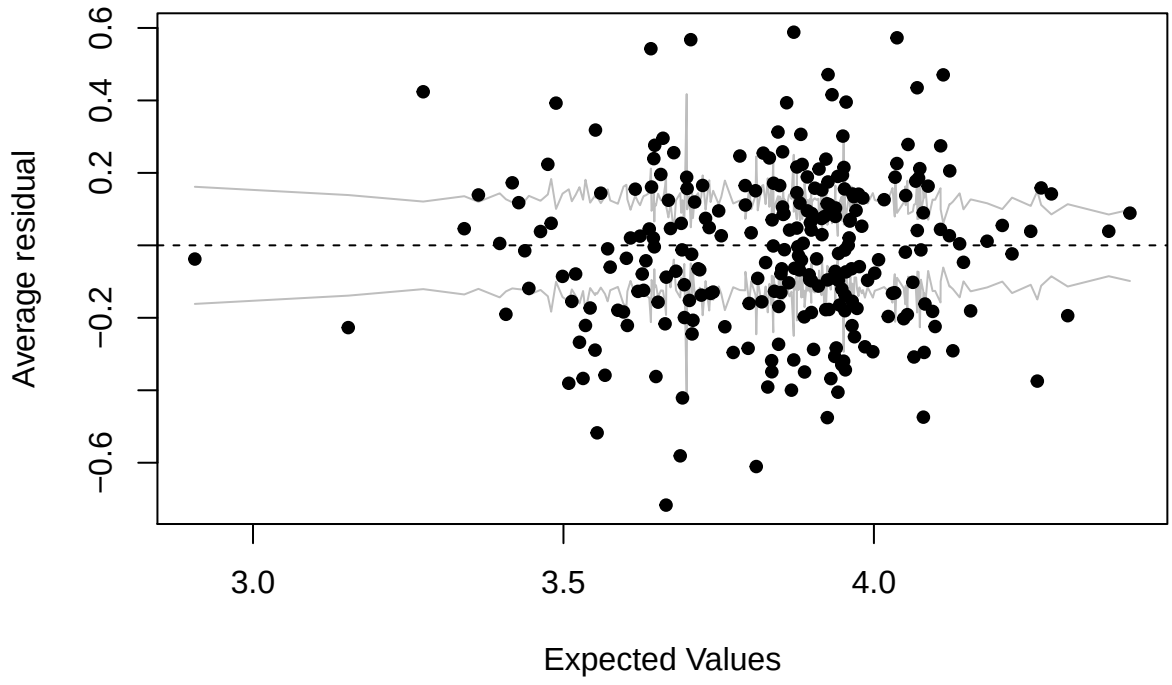
Finally I fit the model only with data of Southeast Asian restaurants.

```
## 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
```

ıre4.6: Binned residual plot of fitting rated stars of Southeast Asian res



4.4 Check and comparethe coefficients table

	Chinese	Japanese	Southeast Asian	Korean
Intercept	3.4604814	3.6375503	3.7802886	4.3633745

	Chinese	Japanese	Southeast Asian	Korean
restaurant_WiFino	-0.0742453	-0.1252521	-0.0001770	-0.3159329
restaurant_WiFipaid	0.3910925	-0.3251938	-0.0840379	0.0178399
restaurant_price_range2	-0.0762948	-0.0172033	0.0168487	-0.2816079
restaurant_price_range3	0.2411818	0.1966469	0.2786113	-0.7230187
restaurant_price_range4	0.2925457	0.1249455	0.5790789	NA
garage_parking true	0.0624246	-0.1297917	-0.2807717	-0.6903903
street_parking true	0.2367694	0.1368207	0.1555946	-0.1428238
validated_parking true	0.3180269	0.1373965	0.2210321	-0.3301659
lot_parking true	0.1697010	0.0323071	0.1063941	-0.0034391
valet_parking true	-0.0963215	-0.2383541	-0.2055574	0.5747600
restaurant_noise_levelloud	-0.1036612	-0.0814839	0.0140009	-0.1348616
restaurant_noise_levelquiet	0.0704491	0.1824398	0.0119266	0.0229342
restaurant_noise_levelvery_loud	-0.8383866	0.1696895	-0.7398966	-0.5233938
restaurant_TV1	0.0810256	-0.0488169	-0.0256351	-0.1244898
restaurant_outdoor_seating1	-0.1163634	-0.0570491	-0.0037261	0.0342456

4.5 Calculate 95% confidence interval and check the statistical significance

Figure4.7: Point estimates and 95% confidence interval of coefficients of model for Chinese restaurants

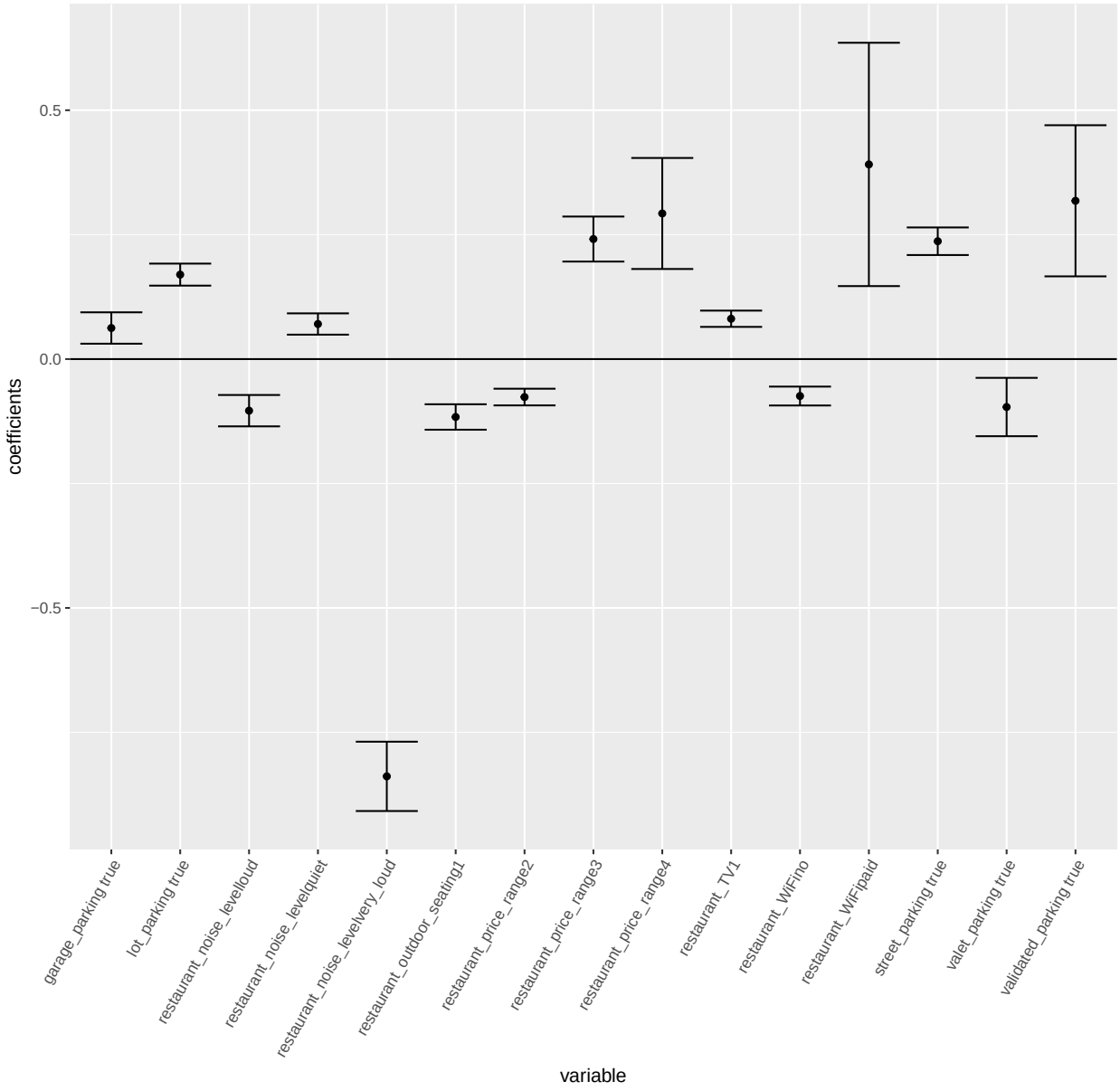


Figure4.8: Point estimates and 95% confidence interval of coefficients of model for Japanese restaurants

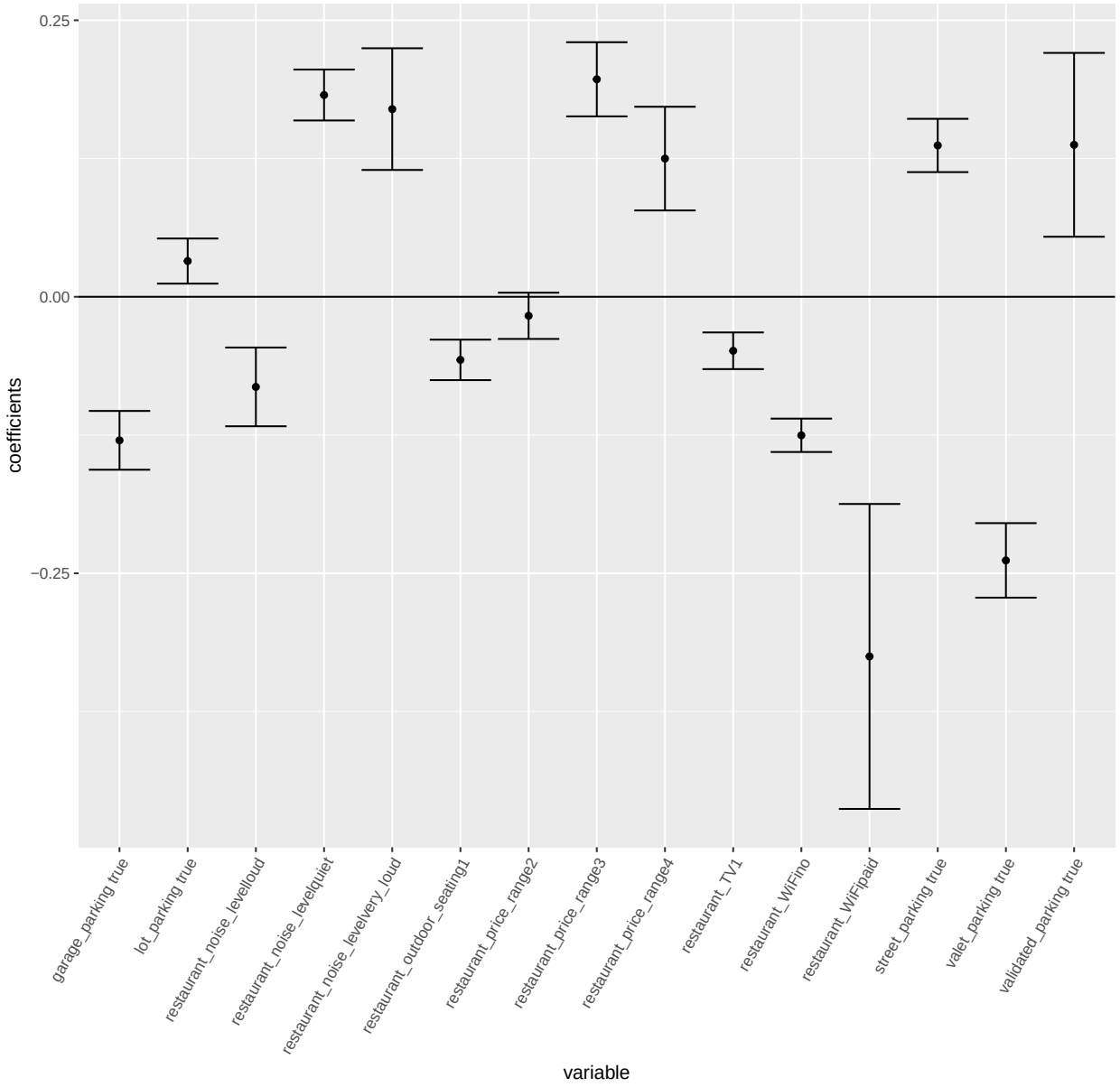


Figure4.9: Point estimates and 95% confidence interval of coefficients of model for Korean restaurants

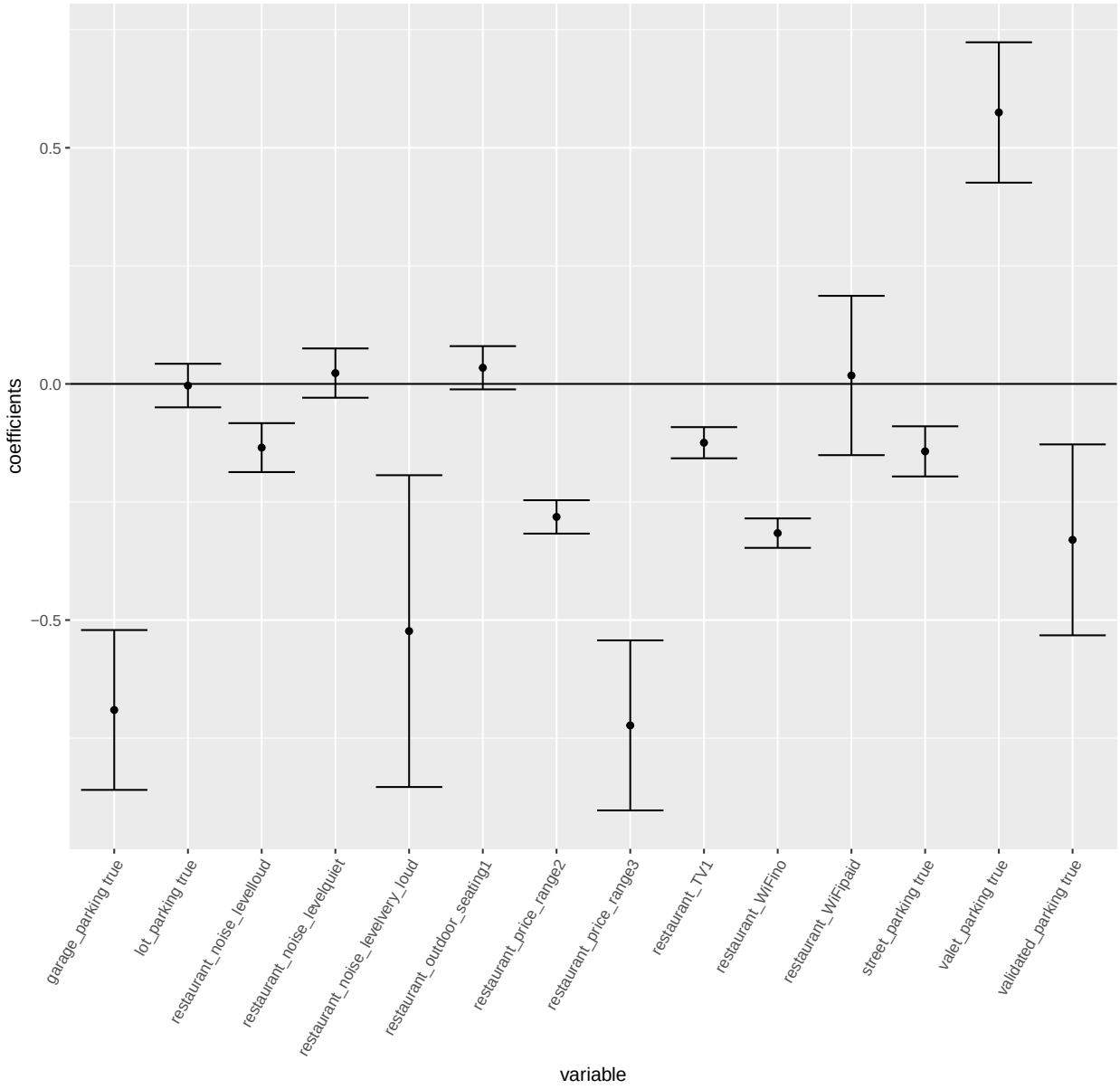
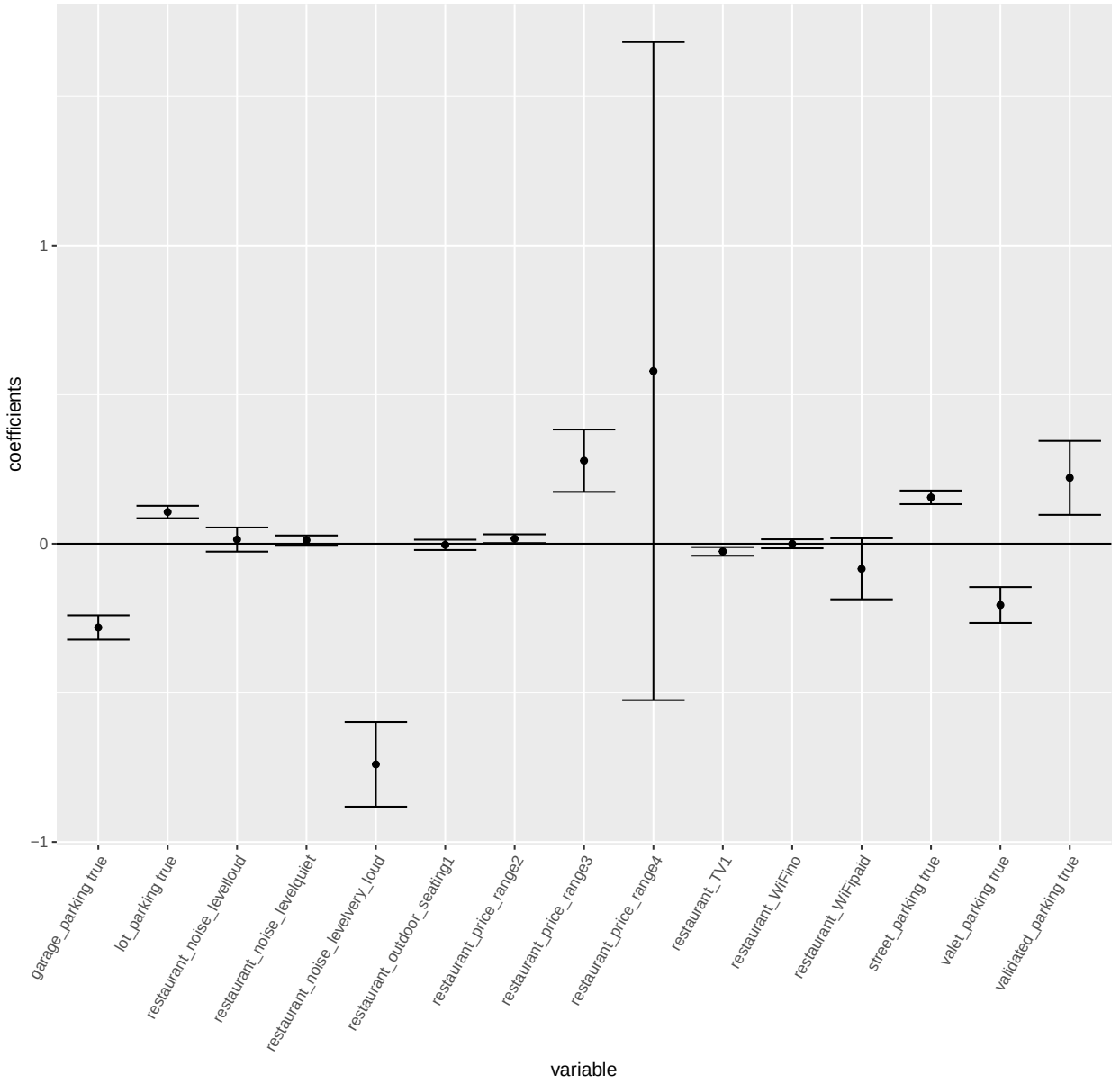


Figure4.10: Point estimates and 95% confidence interval of coefficients of model for Southeast Asian restau



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