application part

Shanshan Song
Ziwen Tang

9 February 2019

1 Background

Suppose we have a dataset that looks at the bounce rates of users of a website with cooking recipes. A bounce rate is a measure of how quickly someone leaves a website, e.g. the number of seconds after which a user first accesses a webpage from the website and then leaves. Most websites want individuals to stay on their websites for a long time as these individuals are more likely to read another article, buy one of their products, click on some of the sponsored links etc. As is said above, it can be useful to understand why some users leave the website quicker than others. The purpose of our study is to work out if younger individuals are more likely to leave the website quicker. See https://www.kaggle.com/ojwatson/mixed-models/notebook for more details.

2 Methodology

2.1 Data Collection

To investigate the bounce rate of the website, three locations were chosen in eight counties in Germany, and members of the public of all ages were requested to fill in a questionnaire. In the questionnaire we asked them to use our search engine to check something they want to eat this evening. Our website was listed in the search engine first, and other similar websites were also included. We recorded the bounce time of surfers who clicked on our website, as well as their age, county and location.

Before we continue we want to standardize the independent explanatory variables by scaling them.

This makes sure that any estimated coefficient from our regression model later on are all on the same scale. So in our case, age would be scaled, thus we have a new variable called age_scaled, which is the age scaled to have zero mean and unit variance. See the listing 1 of dataset sample after scaling in Appendix.

2.2 Data Analysis

Suppose the linear mixed model is constructed as follows

$$bounce\_time = \beta_0 + \beta_1 age\_scaled + b_{c0} + b_{c1} age\_scaled + \epsilon_c \tag{1}$$

In this model, we treat **age_scaled**, which we are interested in, as **fixed effect**, county and location as random effects. To make the model clear and easier to understand, only **county** as **random effect** is considered here. Assumptions about the random effect and error stay the same as before. To fit our model, we mainly use packages "lme4", "arm" and "pbkrtest" to do data analysis. In package "lme4", we apply "lmer" function to fit linear mixed model; in package "arm", "display" function gives a clean printout of regression objects; in package "pbkrtest", "KRmodcomp" function is used to calculate p-value with Kenward and Roger method.

3 Results

3.1 Model Fitted with Different Methods

Fitting the non-multilevel model with OLS and ML:

From listing 2 and 3, model under estimation contains only fixed effect age_scaled, OLS regression and ML regression unsurprisingly give us the same estimated coefficients and standard errors. The fitting value R-Squared is only 0.15, which reminds us to consider a more appropriate model.

Fitting the multilevel model containing fixed effects and only random intercept:

Since we got a low R-Squared value in simple linear model, it's reasonable to add a random intercept in the model. From listing 4, The estimated coefficient of age_scaled is greatly smaller than in ML estimation, the reason is probably that adding a random intercept which explains a large portion of the bounce_time weakens the importance of fixed slope.

Fitting the multilevel model containing fixed effects and only random slope:

Compared to the case above, as can be seen in listing 5, the estimated coefficient of age_scaled in our case is relatively closer to that in ML estimation. That the residual standard deviation is larger than in case above illustrates that mixed model is better fitted with only random intercept than with only random slope.

Fitting the multilevel model containing fixed effects and random effects: The residual standard deviation here in listing 6 is slightly smaller than in case with only random intercept, which possibly implies better fitting.

3.2 Test of Fixed Effects

From the inference part, we know that we will use likelihood ratio test and revised F-test, that is Kenward and Roger, to test fixed effects. There are several R functions which can be used for LRT, in listing 7 and 8 we apply function drop() and anova respectively to test if the coefficient of age_scaled is zero. The p-values from both tests are the same, 0.77. This is larger than the common cut off α level of 0.05. Since this is the lowest we would expect the p-value to be, we have determined that the coefficient of age_scaled is not significant.

To apply Kenward and Roger in our model, we use function KRmodcomp() in pbkrtest package which is specialized for kenward and Roger method. As we can see from listing 9, the coefficient of age_scaled equal zero is being tested and the p-value is 0.8037 which is a bit larger than in LRT, and this gives us the same conclusion that the coefficient of age_scaled is not significant.

3.3 Test of Random Effects

Since we have two random parts-random intercept and random slope in our supposed model, we want to test if the variance parameter of random intercept and random slope equal zero. The riexamp model has only one random effect, random intercept, we want to test if random slope is needed using LRT.

```
rs = -2 * logLik(riexamp) + 2 * logLik(frexamp)#test for random slope
pchisq(as.numeric(rs), df=1, lower.tail=FALSE)
```

The results of the above commands are shown below.

```
[1] 0.0810326
```

The p-value of 0.081 is a bit larger than 0.05, so the variance of random slope is not significantly different from zero, which shows the random slope due to different counties is not significant in our mixed model.

The rsexamp model has only one random effect, random slope, we want to test if random intercept is needed using LRT.

```
ri = -2 * logLik(rsexamp) + 2 * logLik(frexamp)#test for random intercept
pchisq(as.numeric(ri), df=1, lower.tail=FALSE)
```

The results of the above commands are shown below.

```
[1] 4.541321e-62
```

The p-value of 4.541321e-62 is nearly the same as zero, so the variance of random intercept is significantly different from zero, which shows the random intercept due to different counties is significant in our mixed model and we should include random intercept when considering the model.

4 Conclusions

From the result we already analyzed above, we get basically two main conclusions:

1, There is no significant relationship between bounce_time and age, if there exists a very weak relationship between these two, then it should be weakly positive.

2, Counties where individuals come from greatly influence the bounce_time. Namely, individuals from the same county share the same feature regarding how long they will stay in the website.

5 Appendix

Listing 1: dataset sample

```
   bounce_time  age  county  location  age_scaled
1     165.5485   16   devon         a   -1.512654
2     167.5593   34   devon         a   -0.722871
3     165.8830    6   devon         a   -1.951423
4     167.6855   19   devon         a   -1.381024
5     169.9597   34   devon         a   -0.722871
6     168.6887   47   devon         a   -0.152472
```

Listing 2: OLS regression

```
lm(formula = bounce_time ~ age_scaled, data = lmm.data)
            coef.est coef.se
(Intercept) 201.32     0.68
age_scaled    6.28     0.68
---
n = 480, k = 2
residual sd = 14.96, R-Squared = 0.15
```

Listing 3: ML regression

```
glm(formula = bounce_time ~ age_scaled, data = lmm.data)
            coef.est coef.se
(Intercept) 201.32     0.68
age_scaled    6.28     0.68
---
  n = 480, k = 2
  residual deviance = 106970.5, null deviance = 125898.5 (difference = 18928.0)
  overdispersion parameter = 223.8
  residual sd is sqrt(overdispersion) = 14.96
```

Listing 4: Regression with fixed effects and only random intercept

```
Linear mixed model fit by REML ['lmerMod']
Formula: bounce_time ~ age_scaled + (1 | county)
   Data: lmm.data

REML criterion at convergence: 3466.1

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.45717 -0.75415 -0.06246  0.72526  2.56690

Random effects:
 Groups   Name        Variance Std.Dev.
 county   (Intercept) 213.03   14.595
 Residual              74.73    8.645
Number of obs: 480, groups:  county, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept) 201.3165     5.1753  38.899
age_scaled    0.1355     0.6107   0.222

Correlation of Fixed Effects:
           (Intr)
age_scaled 0.000
```

Listing 5: Regression with fixed effects and only random slope

```
Linear mixed model fit by REML ['lmerMod']
Formula: bounce_time ~ age_scaled + (-1 + age_scaled | county)
   Data: lmm.data

REML criterion at convergence: 3739.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2657 -0.6435  0.0397  0.6483  2.7189

Random effects:
 Groups   Name       Variance Std.Dev.
 county   age_scaled 110.8    10.53
 Residual            134.9    11.62
Number of obs: 480, groups:  county, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept) 205.6057     0.7562 271.875
age_scaled    4.5997     3.7648   1.222

Correlation of Fixed Effects:
           (Intr)
age_scaled 0.005
```

Listing 6: Regression with fixed effects and random effects

```
Linear mixed model fit by REML ['lmerMod']
Formula: bounce_time ~ age_scaled + (1 + age_scaled | county)
   Data: lmm.data

REML criterion at convergence: 3463

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.47807 -0.77894 -0.05381  0.72267  2.57503

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 county   (Intercept) 198.73   14.097
          age_scaled    1.99    1.411   -0.85
 Residual              74.13    8.610
Number of obs: 480, groups:  county, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept) 201.9044     5.0030  40.357
age_scaled    0.2045     0.7832   0.261

Correlation of Fixed Effects:
           (Intr)
age_scaled -0.540
```

Listing 7: Likelihood ratio test for fixed effect using drop()

```
Single term deletions

Model:
bounce_time ~ age_scaled + (1 + age_scaled | county)
           Df    AIC      LRT Pr(Chi)
<none>        3480.9
age_scaled  1 3479.0 0.085499    0.77
```

Listing 8: Likelihood ratio test for fixed effect using anova

```
Data: lmm.data
Models:
mmDx1: bounce_time ~ 1 + (1 + age_scaled | county)
mmMLE: bounce_time ~ age_scaled + (1 + age_scaled | county)
      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
mmDx1  5 3479.0 3499.9 -1734.5   3469.0
mmMLE  6 3480.9 3506.0 -1734.5   3468.9 0.0855      1       0.77
```

Listing 9: Kenward and Roger approximation for fixed effect

```
F-test with Kenward-Roger approximation; computing time: 1.15 sec.
large : bounce_time ~ age_scaled + (1 + age_scaled | county)
small : bounce_time ~ 1 + (1 + age_scaled | county)
        stat    ndf    ddf F.scaling p.value
Ftest 0.0668 1.0000 6.8772         1  0.8037
```