

Mixed Model Analysis For a Website Bounce Rate

McBeth Ahorator & Noah Gblonyah

December 7, 2022

Introduction

Bounce rate is a word used in internet marketing and business to describe how quickly someone leaves a website e.g. the number of seconds after which a user first accesses a webpage from a website and then leaves. The rate denotes the proportion of visitors that arrive to the site and then leave versus those who come and see numerous pages. The bounce rate simply assists a firm in better understanding how “sticky” its website is — how effectively it maintains users’ interest. Lower bounce rates are more successful in attracting visitors and keeping users on the site to read numerous pages. When designing a website that would entice people to visit more than one page, there are several factors to consider. The three most important approaches to keep bounce rates low are to ensure that the website successfully delivers content, is aesthetically beautiful, and is user-friendly. Most webpages want users to remain on their websites for an extended period of time since they are more inclined to read another article, purchase one of their items, click on some of the sponsored links, and so on. As a result, understanding why certain users leave the page faster than others might be beneficial.

To explore the website’s bounce rate, three sites in eight (8) counties in England were chosen, and members of the general public of all ages completed a survey. Participants were invited to utilize our restaurant website’s search engine to search for food to eat in the evening. The restaurant’s website, as well as other websites, were listed by the search engine. Users that visited the restaurant’s website were then timed and their bounce rate was reported. This project seeks to ascertain if younger individuals are more likely to leave the website quicker.

Data Overview

This dataset consists of 480 observations and contains information on bounce time, age, sex, county and location of users in England. This dataset to be used was extracted from Kaggle.

Variables	Description
bounce_time	The number of seconds a user spends on the website
age	the age in years of the user
Sex	Gender (0=male, 1=female)
county	8 counties in England; Cheshire, Cumbria, Devon, Dorset, Essex, Kent, London, Norfolk
location	3 locations; a, b and c

We have data collected from three (3) in eight (8) counties in England. Each county has sixty (60) users visiting the website. The counties with the minimal bounce rates are London with 162.33 seconds, and Devon with 164.42 seconds. The maximum bounce rate is recorded in Cheshire with 236.36 seconds and Cumbria 229.91 seconds. There appears not to be any missing data or typos.

Table 2: Summary statistics on the bounce time by county

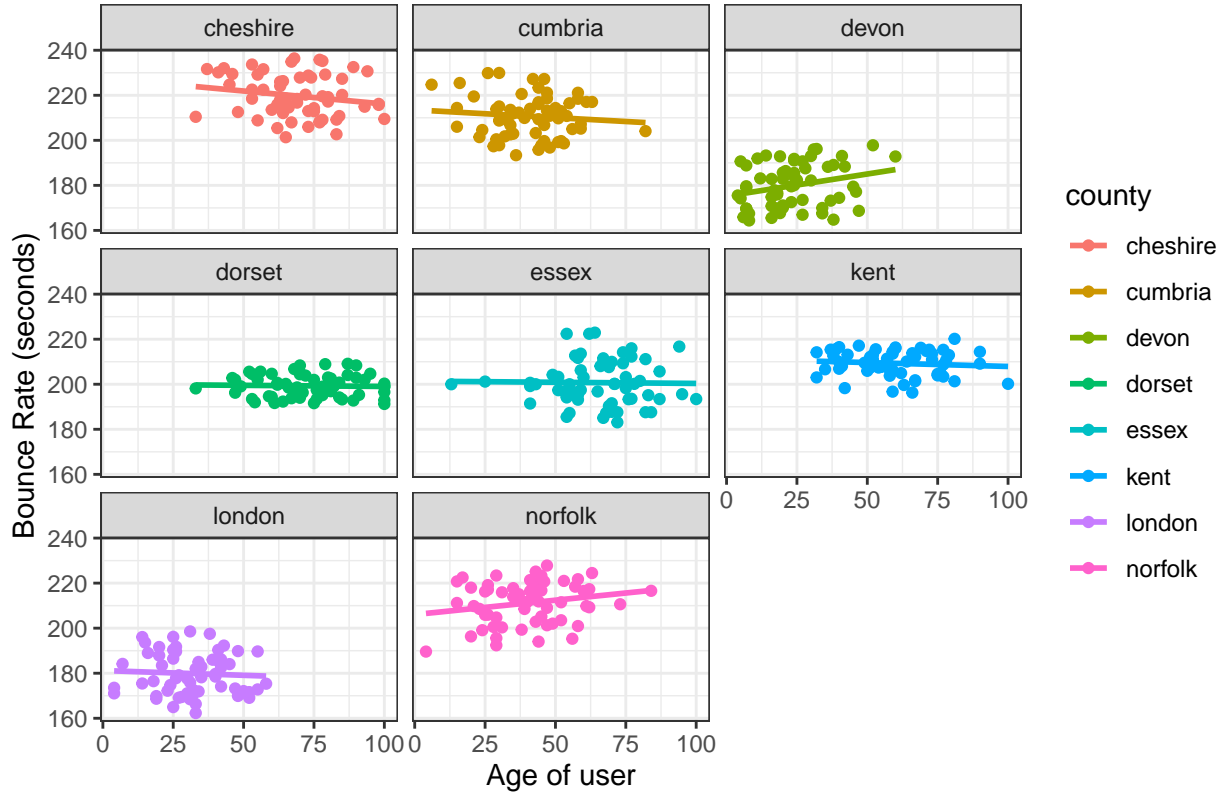
county	min	Q1	median	Q3	max	mean	sd	n
cheshire	201.3748	212.8307	218.4270	228.0927	236.3625	219.7284	9.348592	60
cumbria	193.3905	203.1536	210.8615	216.9494	229.9114	210.6880	9.265984	60
devon	164.4163	170.7862	179.4992	188.4725	197.7972	179.8769	9.679149	60
dorset	191.2489	195.0444	198.9579	202.9248	209.0729	199.2856	4.955419	60
essex	183.0902	193.4214	200.0066	208.3685	222.9063	200.7198	10.016569	60
kent	196.2828	205.7073	209.7167	213.8786	220.1563	209.2756	5.622894	60
london	162.3266	172.0287	178.3485	186.8189	198.5600	179.8393	9.266892	60
norfolk	189.6057	204.4329	211.7890	218.0982	227.8138	211.1180	9.342629	60

Table 3: Summary statistics on the bounce time by county

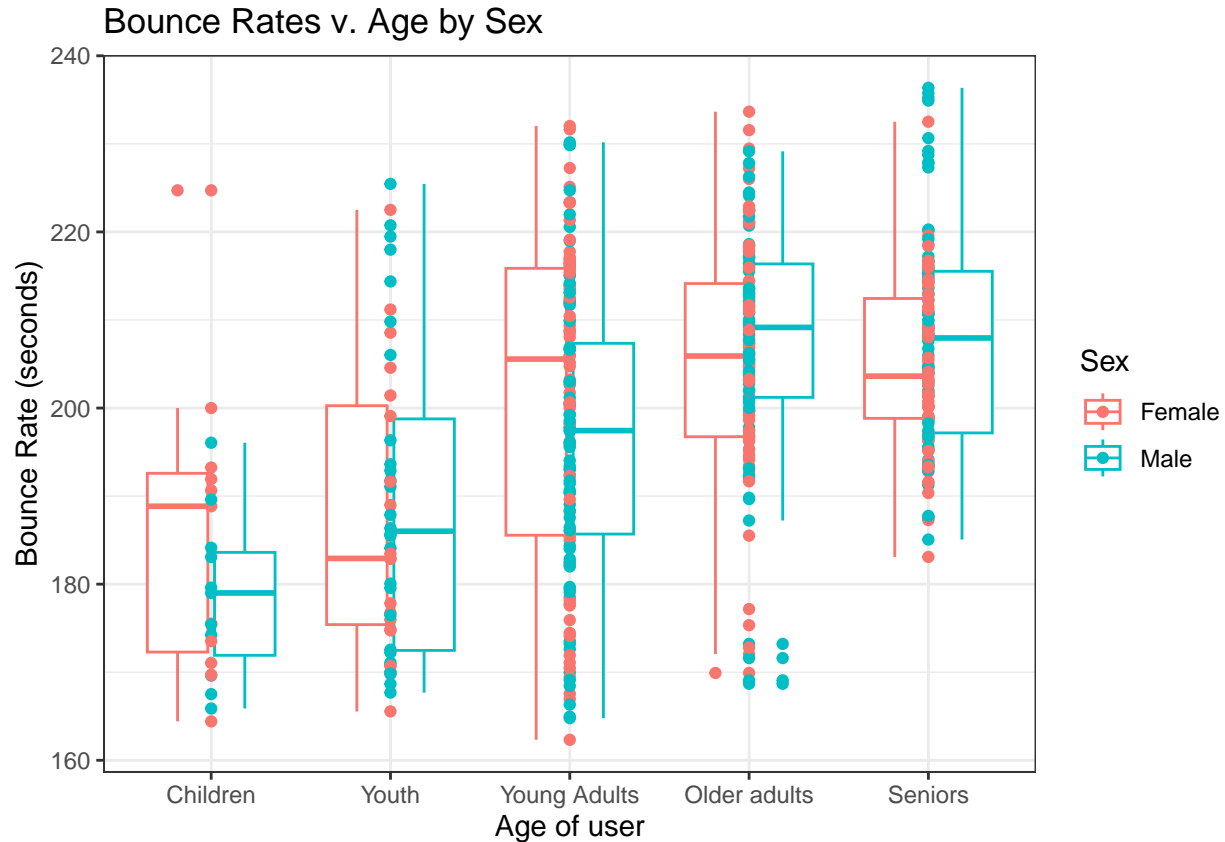
location	min	Q1	median	Q3	max	mean	sd	n
a	162.3266	182.8198	198.4023	205.4664	218.5710	194.1747	15.45886	160
b	169.0763	191.4188	204.6230	213.2250	232.0245	201.2327	15.46286	160
c	183.4594	196.1186	209.0080	220.8072	236.3625	208.5420	14.48687	160

The study had one-sixty (160) users visiting the website in each of the three (3) locations chosen for the study (a, b, c). The minimum bounce rates for locations a, b, and c are 162.33, 169.08, and 183.46 respectively. The maximum bounce rates for locations a, b, and c are 218.57, 232.02, and 236.36 respectively. The highest mean bounce rate is 208.54 seconds recorded in location c, and lowest mean bounce rate is 218.57 seconds recorded in location a.

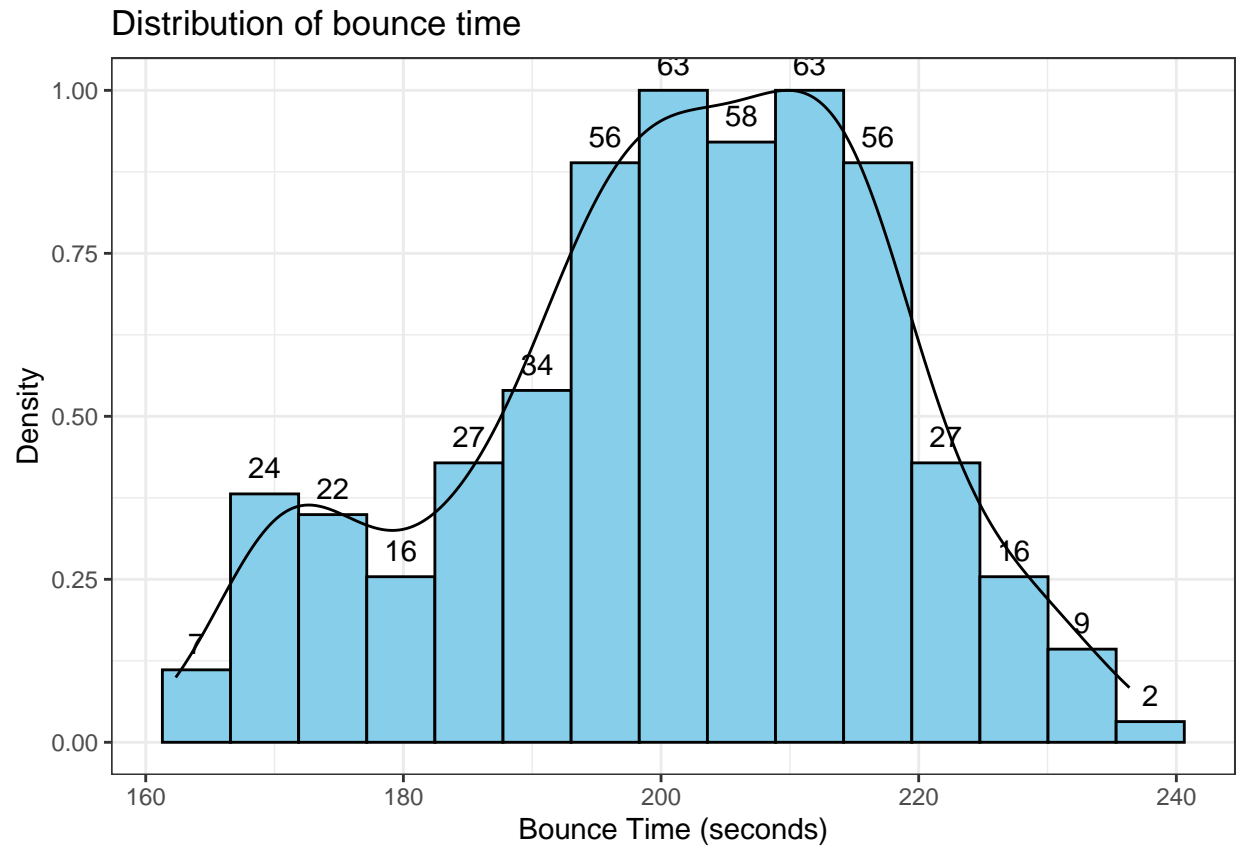
Bounce Rates v. Age by County



We are looking at the Bounce rate for different age groups across the counties. From the figure, we can see that age has some relationship with the amount of time an individual spends on the website. For people in Devon, young individuals tend to spend less time as compared to adults. When we look at this relationship based on the specific counties, we can see that this is not always the case. People from a specific county tend to behave similarly when it comes to their bounce rate for the respective age groups. People in Norfolk spend more time on the website than those in London on average. Also, as the youth in Devon are less interested in the site than adults, the youth or young ones from Cumbria tend to have a higher bounce rate as compared to the adults for the website.



We can see from the boxplots that, children (age < 14) spend have the least bounce rate, and older adults (age between 45 and 64), seniors (age > 64) have the highest average bounce rates. Youth tend to also spend less time on the restaurant website. Among children and young adults, females tend to spend more time on the website compared to male. This varies for Youth, Older adults, and seniors where male spend more time on the website, on average.

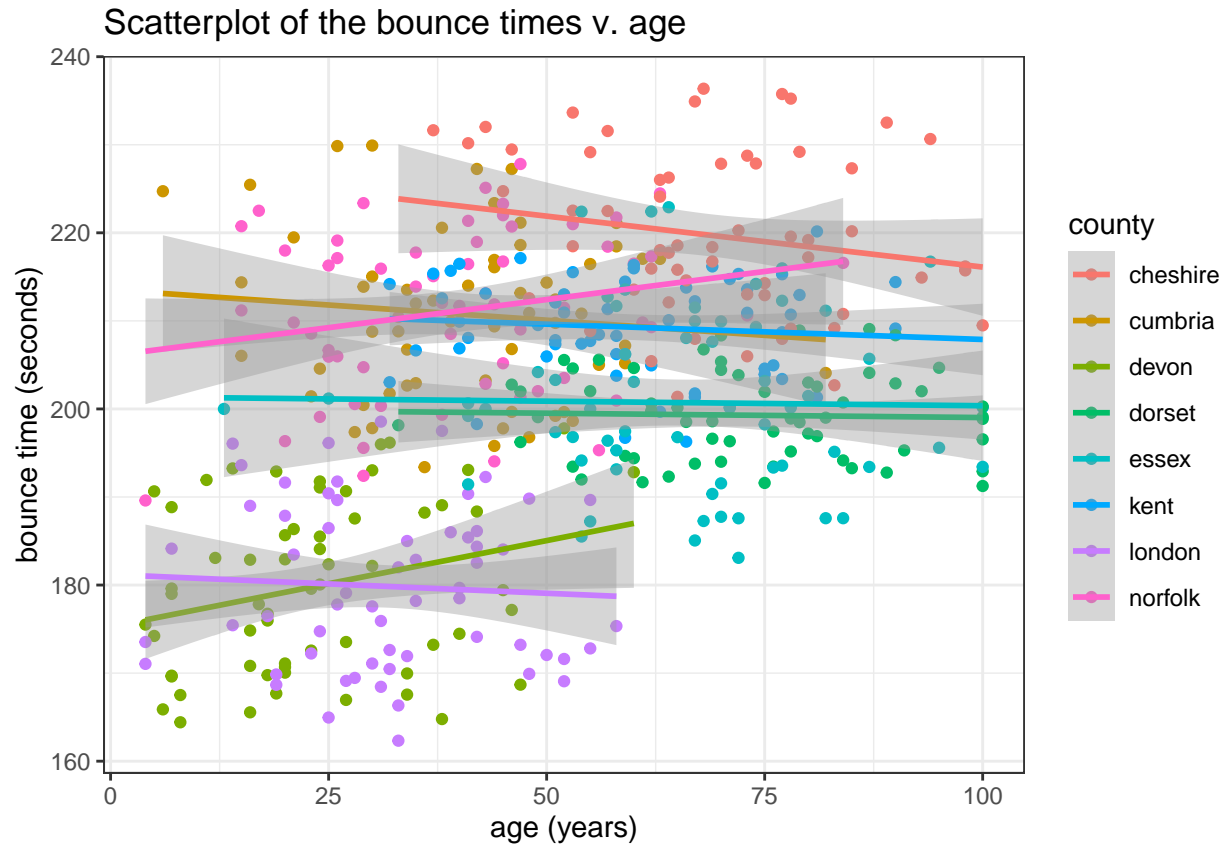


The distribution of the bounce rates is approximately normally distributed with a few spikes between 160 and 175 seconds. About 90% of the data lies between 180 to 230 seconds.

Statistical Modeling Procedure

Our main research question is to know if the age of an individual affects the time in seconds they spend on the website. We will consider the bounce rate as our response variable and the location, county, age, and sex of the individuals as the explanatory variables. The model to answer this research question will consider the location and county as random effects because our exploratory data analysis gives us a clue on how they should be treated since the bounce rates vary within the groups for these variables. We will first consider a combination of models which includes all fixed effects and possible interactions and select the best model using the anova function which computes deviance tables for the fitted model objects and tests the models against one another in the order specified for model selection. After we settle on a fixed effect model, we will also use the AIC function for the selection of our best possible random effect structure that should be included in the model.

1. Simple Linear Regression



According to the scatterplot above, older users seem to spend more time on the website. To go further, we want to explore if bounce time is affected by age. A simple linear regression would be used to estimate these coefficients. The model can be specified by

$$y_i = \beta_0 + \beta_1 x_{age} + \beta_2 I_{sex=1} + \beta_3 I_{sex=1} x_{age} + \epsilon_i$$

where:

y_i represents the bounce time for an individual $\beta_j, j = 0, 1, 2, 3$ represents the parameter estimates x_{age} represents the age of an individual $I_{sex=1}$ represents the sex of the individual with 1 representing females $\epsilon_i \sim N(0, \sigma^2)$ represents the error in the model.

2. Hierarchical model with Fixed Slope and Random Intercept

Due to the nature of the data, a mixed effects model is ideal here as it will allow us to both use all the data we have and better account for the correlations within data coming from the same counties and locations. The proposed model is given by:

$$\begin{aligned} y_i &\sim N(\alpha_{j[i]} + \beta_1 x_{age} + \beta_2 I_{sex=1} + \beta_3 I_{sex=1} x_{age}, \sigma_y^2) \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha^2) \end{aligned}$$

where

- $i = 1, 2, \dots, 480$ are the number of users in the data
- $j = 1, 2, \dots, 8$ are the counties
- $j[i]$ denotes the counties for the i th observation
- y_i is the average bounce time of the website visit i in county j .
- α_j represents the intercept for county j .
- β_j represents the overall slope across counties j .
- x_{age} represents the age of a user i in counties j .
- μ_α overall mean for the intercept of counties j .
- σ_y^2 is the overall error in the data.
- σ_α^2 is the variance of the counties means that is, the distance from μ_α for the county means.

3. Mixed Effects model with Nested Random Intercepts

One way to incorporate the impact of county on age is to vary intercept on the location variable as well. This model can be specified below as:

$$\begin{aligned} y_i &\sim N(\alpha_{j[i]} + \beta_1 x_{age} + \beta_2 I_{sex=1} + \beta_3 I_{sex=1} x_{age}, \sigma_y^2) \\ \alpha_j &\sim N(\mu_\alpha, \sigma_\alpha^2) \end{aligned}$$

where

- $i = 1, 2, \dots, 480$ are the number of cupcakes in the data
- $j = 1, 2, \dots, 8$ are the locations
- $j[i]$ denotes the location for the i th observation
- y_i is the average bounce time of the website visit i in location j .
- $\alpha_{j[i]}$ represents the intercept for location in county j controlling for other variables.
- $\beta_{j[i]}$ represents the slope for location in county j after controlling for other variables
- x_{age} represents the age of a user i in location j .
- μ_α overall mean for the intercept of location j .
- σ_y^2 is the overall error in the data.
- σ_α^2 is the variance of the location means that is, the distance from μ_α for the location means.

Model Comparison and Defense of Model Choice

To assess how well a model explains the data, two widely used methods were adopted:

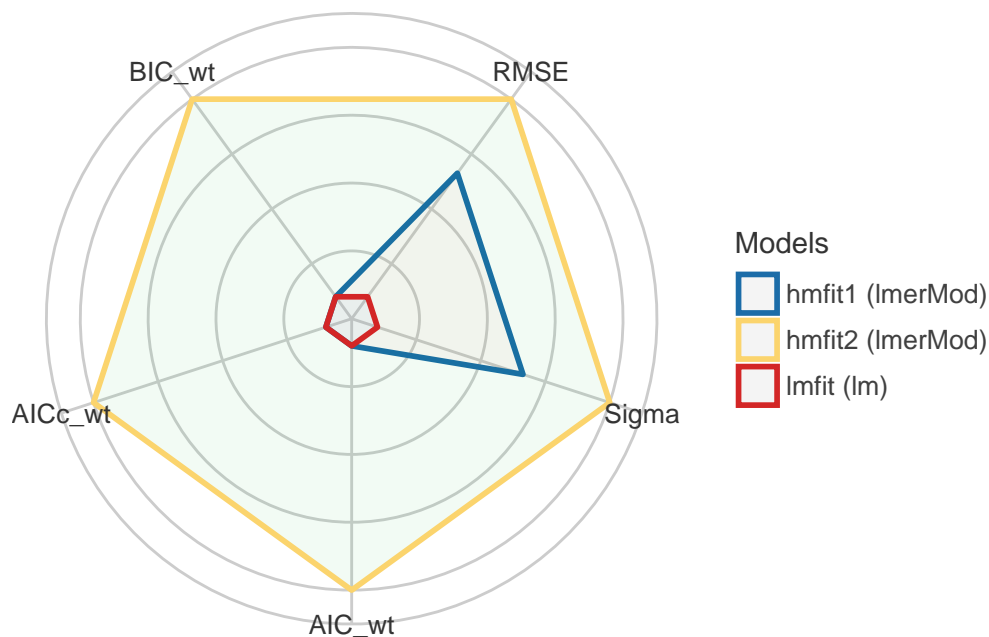
- Akaike's information criterion (AIC): Akaike (1998) proposed AIC as a measure of model quality which can be expressed as $AIC = -2\log L(\hat{\theta}) + 2k$, where θ is the set of parameters of the model, $L(\hat{\theta})$ is the likelihood of the proposed model given the data when evaluated at the maximum likelihood estimate of θ , and k is the number of estimated parameters in the proposed model.
- Bayesian information criteria (BIC): Unlike the AIC, the BIC penalizes free parameters more strongly and is computed according to Schwarz (1978) as $BIC = -2\log L(\hat{\theta}) + k\log(n)$

Name	Model	AIC	BIC	logLik
lmfit	lm	3965.3	3982.0	-1978.6
hmfit1	lmerMod	3481.7	3502.6	-1735.9
hmfit2	lmerMod	3024.2	3049.3	-1506.7

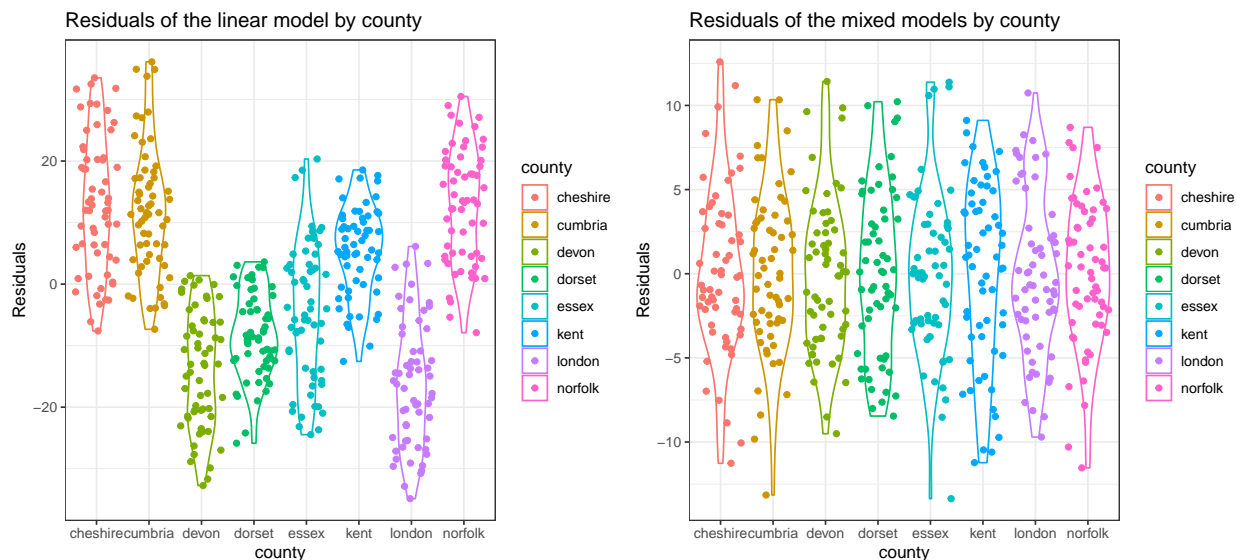
The table values shows the performance measures (AIC, BIC) of the three models proposed namely, the simple linear model (lmfit), the mixed effect random intercept with fixed slope model(hmfit1) and the nested mixed effect random intercept model (hmfit2). The `lmfit` model recorded an AIC value of 3965.3 and BIC value of 3982. The `hmfit1` model recorded an AIC value of 3481.7 and a BIC value of 3502.6. Finally, the `hmfit2` model recorded an AIC value of 3024.2 and a BIC value of 3049.3. The nested mixed effect model (hmfit2) outperformed the other two models since it had the lowest recorded value for AIC and BIC. Hence on the basis of these performance measures `hmfit2` is considered the best model for this data.

```
## Data: bounce
## Models:
## lmfit: bounce_time ~ age * Sex
## hmfit1: bounce_time ~ age * Sex + (1 | county)
## hmfit2: bounce_time ~ age * Sex + (1 | location/county)
##      npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## lmfit      5 3965.3 3986.2 -1977.7   3955.3
## hmfit1     6 3483.7 3508.7 -1735.8   3471.7 483.66  1 < 2.2e-16 ***
## hmfit2     7 3025.7 3055.0 -1505.9   3011.7 459.92  1 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Comparison of Model Indices



This plot creates a “spiderweb” plot, where the different indices are normalized and larger values indicate better model performance. Hence, points closer to the center indicate worse fit indices.



One of the other key assumptions of the simple linear model is that the observations of our data are independent of the other data as and the constant variance assumption. When we collected our data we were doing it in 8 different counties and in 3 locations within each county. So we could check this by comparing the bounce times for each county against the residuals. The “Residuals vs Fitted” panel in the left panel displays the residuals on the y-axis and the counties on the x-axis for the `lm` model. Generally, the residuals seem to differ greatly between groups but look somewhat similar within groups. clearly there is substantial

grouping of each county’s residual. So we can definitely say that the residuals in our data are not constant but rather heteroscedastic. In summary, the `lm` model assumes that the errors are independent when in fact, they are not and thus it is inappropriate to use a linear model for this data. The right panel plot is the residual vs fitted plot for the `mem` model. These residuals seems to be centered around 0 on average with the residuals much better distributed, illustrating that the random effect, county, has accounted for the correlation structure in the model. Moreover, with the `mem` model, the right standard errors are correctly obtained which accurately accounts for uncertainty in prediction and estimation.

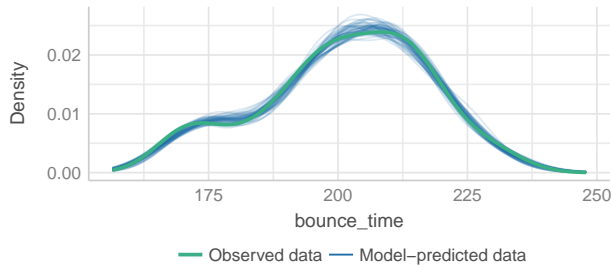
Results and Discussion

Model Assumptions

The linear regression model assumes that there exists a linear relationship between the independent variables and the dependent variable, and that the errors of the model are independent. It also assumes that our model residuals have a constant variance at every level and that the residuals of the model are also normally distributed. The independence of data is one of the GLMs’ assumptions. As a result, this assumption implies that the underlying research design is entirely randomized. However, according to Robinson et al. (2004), in practice, this assumption is commonly violated. A common example is the collection of longitudinal data, in which an observational unit is examined across time, or in the case of our data for this study, where the data has a clustered structure. The resulting data is correlated, which violates the independence assumption. According to Robinson et al. (2004), hierarchical models (HM) extend the generalized linear model (GLM) to account for correlations in random effects. In summary, the GLM assumes that the errors are independent when in fact, they are not. This produces underestimated standard errors of the estimates leading to inaccurate prediction. However, with the HM, the right standard errors are correctly obtained which accurately accounts for uncertainty in prediction and estimation.

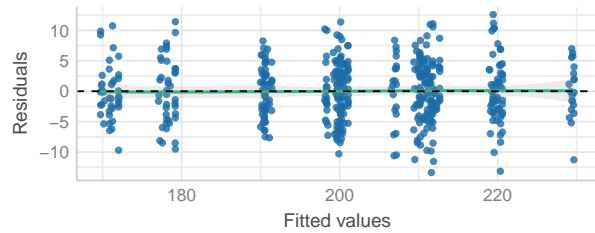
Posterior Predictive Check

Model-predicted lines should resemble observed data line



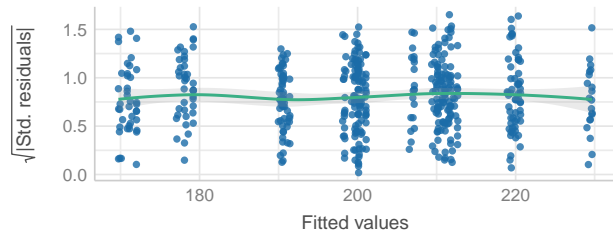
Linearity

Reference line should be flat and horizontal



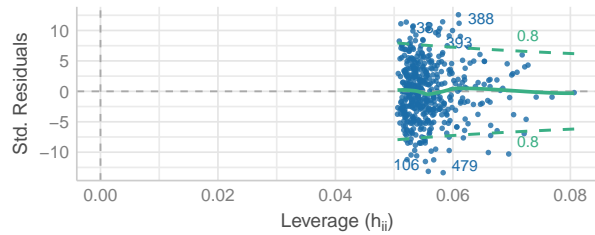
Homogeneity of Variance

Reference line should be flat and horizontal



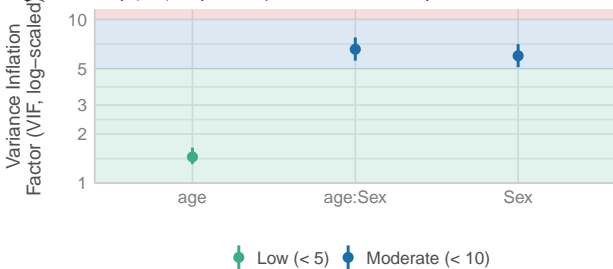
Influential Observations

Points should be inside the contour lines



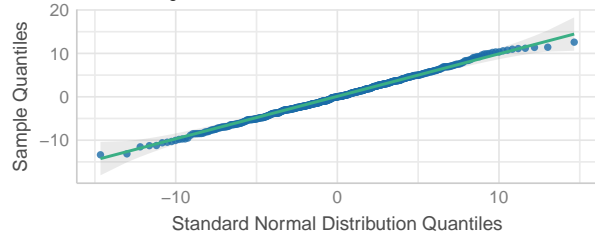
Collinearity

High collinearity (VIF) may inflate parameter uncertainty



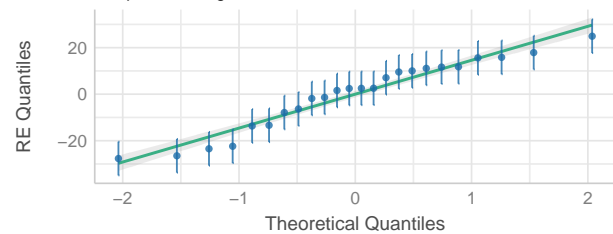
Normality of Residuals

Dots should fall along the line



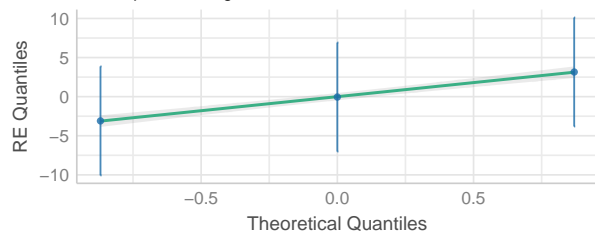
Normality of Random Effects (county:location)

Dots should be plotted along the line



Normality of Random Effects (location)

Dots should be plotted along the line



The plot shows the various diagnostic plots associated for checking the assumptions of a mixed model. These assumptions are similar to that of a simple or generalized linear model. The residuals vs fitted plot in the top right corner shows the spread of the residuals in the 8 counties. It can be observed that the residuals seem to be spread evenly across the counties which shows that the assumption of linearity is met as well as the homogeneity of variance. There also appears not to be any influential points based on the leverage plot. The first plot on the third row shows that there is little to moderate collinearity issues with respect to our variables. The QQ plot illustrates no severe violations of the the normality assumption as almost all points lie on the line. similarly, an additional assumption of the mixed effects model over the linear model is that the random effects should be normally distributed. The last two plots show that there is no severe violation of this assumption and hence the random effects are distributed normally.

Results of Nested Mixed Effects Model

```
## Linear mixed model fit by REML ['lmerMod']
```

```

## Formula: bounce_time ~ age * Sex + (1 | location/county)
## Data: bounce
##
## REML criterion at convergence: 3019.4
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.7276 -0.6202  0.0022  0.6632  2.5695
##
## Random effects:
##   Groups             Name             Variance Std.Dev.
## county:location (Intercept) 231.21    15.206
## location        (Intercept)  22.43     4.737
## Residual                        24.01     4.900
## Number of obs: 480, groups:  county:location, 24; location, 3
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept) 201.66145    4.26151  47.322
## age         -0.00135    0.01925  -0.070
## SexMale     -1.23561    1.12860  -1.095
## age:SexMale  0.01425    0.02033   0.701
##
## Correlation of Fixed Effects:
##              (Intr) age    SexMal
## age          -0.228
## SexMale     -0.138  0.491
## age:SexMale  0.130 -0.555 -0.913

##      (Intercept)          age          SexMale    age:SexMale
## 201.661451980 -0.001349817 -1.235606393  0.014250632

```

From the output, it can be seen that the estimated average bounce time of a user for an average age across all counties and locations is 201.66 seconds with an error margin of 4.3 seconds. The bounce time of a user on the website across all counties and in all locations is estimated to decrease by 0.04 seconds (relatively zero) with an error margin of 0.06 seconds for an additional increase in the age of a user. This can mean the impact of age across counties and locations is negligible. The overall error in the model is estimated to be 6 seconds.

The 95% confidence interval for the fixed effects is (187.5, 222.1) seconds, this means that the overall bounce time for website visit ranges between 187.5 seconds and 222.1 seconds on average. The 95% intervals for the county effects (or deviations from the mean bounce time) are:

Also, the 95% intervals for the location effects (or deviations from the mean bounce time) are:

Table 5: Interval estimates of random effects for various counties

County	lower	upper
cheshire:a	4.4	19.2
cheshire:b	10.5	25.3
cheshire:c	17.5	32.3
cumbria:a	-4.8	10.0
cumbria:b	2.6	17.4
cumbria:c	8.2	23.0
devon:a	-35.1	-20.3
devon:b	-29.8	-15.0
devon:c	-20.8	-6.0
dorset:a	-5.8	9.0
dorset:b	-9.3	5.6
dorset:c	-13.8	1.1
essex:a	-15.3	-0.5
essex:b	-8.8	6.0
essex:c	-0.3	14.5
kent:a	4.2	19.0
kent:b	2.1	17.0
kent:c	-5.0	9.9
london:a	-33.9	-19.1
london:b	-30.9	-16.1
london:c	-21.1	-6.3
norfolk:a	-4.9	9.9
norfolk:b	3.7	18.5
norfolk:c	8.4	23.2

Table 6: Interval estimates of random effects for various locations

Location	lower	upper
a	-10.2	4.0
b	-7.2	7.1
c	-4.0	10.3

The estimated standard error associated with the intercept of counties is 14.3 seconds and the estimated standard error for intercept associated with locations is 12.2 seconds. Based on the output, crucially though, we can see that age does not impact the bounce time after we have controlled for the random variation caused by the county properly and location, i.e. with a random slope and intercept. It can be seen from our output that users from London county spend 24.4 seconds less on the website on average, followed by Devon, with 24.1 seconds less on average. On average, Cheshire county is estimated to spend more time on the website (19 seconds) than any other county. The addition of random effects permits generalizations to the population of England from which subjects were sampled, accounts for differences between subjects, and accounts for within subjects dependency. Hence inferences can be made to new counties and new locations in England.

References

- Waisberg, D., & Kaushik, A. WAISBERG, D. AND KAUSHIK, A.—WEB ANALYTICS: EMPOWERING CUSTOMER CENTRICITY Web Analytics 2.0: Empowering Customer Centricity.
- <https://corporatefinanceinstitute.com/resources/knowledge/other/bounce-rate/>

- A. Kaushik. Excellent analytics tip 11: Measure effectiveness of your web pages. Occam's Razor (blog), May 2007.
- <https://www.kaggle.com/code/ojwatson/mixed-models/>
- Robinson, T. J., Myers, R. H., & Montgomery, D. C. (2004). Analysis considerations in industrial split-plot experiments with non-normal responses. *Journal of Quality Technology*, 36(2), 180-192.
- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199-213). Springer, New York, NY.
- Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.
- Hao Zhu (2021). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. R package version 1.3.4. <https://CRAN.R-project.org/package=kableExtra>
- Yihui Xie (2021). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.34.
- Goodrich B, Gabry J, Ali I & Brilleman S. (2020). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.1 <https://mc-stan.org/rstanarm>.
- Wickham et al., (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686, <https://doi.org/10.21105/joss.01686>
- Taiyun Wei and Viliam Simko (2021). R package 'corrplot': Visualization of a Correlation Matrix (Version 0.90). Available from <https://github.com/taiyun/corrplot>
- Wickham, H., & Wickham, M. H. (2017). Package tidyverse. Easily Install and Load the 'Tidyverse'.
- Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. <https://CRAN.R-project.org/package=ggpubr>
- Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC. ISBN 978-1466561595

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(lme4)
library(tidyverse)
library(mosaic)
library(knitr)
library(tidyverse)
library(arm)
library(lme4)
library(kableExtra)
library(coefplot)
library(see)
library(performance)
library(ggpubr)
library(sjPlot)
library(sjmisc)
library(nlme)

bounce <- read.csv("bounce.csv")
bounce$Sex[bounce$Sex=="1"] = "Female"
bounce$Sex[bounce$Sex=="0"] = "Male"

table1 <- favstats(bounce ~ county, data = bounce)
table1$missing <- NULL
```

```

kable_styling(kable(table1,
caption = " Summary statistics on the bounce time by county"),
latex_options = c("hold_position", "striped" ))
table2 <- favstats(bounce ~ location, data = bounce)
table2$missing <- NULL
kable_styling(kable(table2,
caption = " Summary statistics on the bounce time by county"),
latex_options = c("hold_position", "striped" ))

ggplot(bounce,aes(x = age, y = bounce_time, color = county))+
  geom_point()+ geom_smooth(formula = 'y~x', method = "lm", se = FALSE)+
  facet_wrap(~county)+
  theme_bw()+
  labs(title = "Bounce Rates v. Age by County", x = "Age of user", y = "Bounce Rate (seconds)")
# ggplot(bounce,aes(x = age, y = bounce_time, color = location))+
#   geom_point()+ geom_smooth(formula = 'y~x', method = "lm", se = FALSE)+
#   facet_wrap(~location)+
#   theme_bw()+
#   labs(title = "Bounce Rates v. Age by Location", x = "Age of user", y = "Bounce Rate (seconds)")
bounce$Age.c <-cut(bounce$age, breaks=c(0, 14, 24, 45, 64, 110),
  labels=c("Children", "Youth", "Young Adults", "Older adults", "Seniors"))

ggplot(bounce,aes(x = Age.c, y = bounce_time, color = Sex))+ geom_boxplot()+
  geom_point()+ theme_bw()+
  labs(title = "Bounce Rates v. Age by Sex", x = "Age of user", y = "Bounce Rate (seconds)")
ggplot(bounce, aes(x = bounce_time))+
  geom_histogram(aes(y = ..ncount..), bins = 15, col = 1, fill = "skyblue", center = 0) +
  geom_density(aes(y = ..scaled..)) +
  theme_bw() +
  labs(y = "Density") +
  stat_bin(aes(y = ..ncount.., label = ..count..), bins = 15, geom = "text", vjust = -0.75)+
  labs(title = "Distribution of bounce time",
  x = "Bounce Time (seconds)")
ggplot(bounce, aes(y = bounce_time, x = age, color = county)) +
  geom_point() + theme_bw() + geom_smooth(formula = 'y~x', method = "lm")+
  labs(title = "Scatterplot of the bounce times v. age",
  x = "age (years)",
  y = "bounce time (seconds)")
hmfit1 <- lmer(bounce_time ~ age*Sex + (1 | county), data = bounce)
hmfit2 <- lmer( bounce_time ~ age*Sex + (1| location/county), data = bounce)
lmfit <- lm(bounce_time ~ age*Sex, data = bounce)

fit1 <- lme(bounce_time ~ age + Sex, random = list(~1|county, ~1|location), data = bounce)
fit2 <- lme(bounce_time ~ age + Sex, random = ~1| county, data = bounce)

#anova(lmfit,fit1,fit2)
anova(hmfit2,lmfit,hmfit1)
bounce$resid <- residuals(hmfit2)

f <- ggplot(bounce, aes(x = age, y = resid, col = county, group = location)) +
  geom_line() +
  facet_wrap(~ county)

```

```

f2 <- ggplot(bounce, aes(x = age, y = resid, col = county)) +
  geom_point() +
  geom_smooth(formula = 'y~x', method = "lm")

fit11 <- lmer(bounce_time ~ age + Sex +(1+age|location/county), data = bounce)
an <- anova(hmfit2, fit11)
# anova still favors model without random slope
modelcheck <- compare_performance(lmfit, hmfit1, hmfit2, rank = TRUE)
plot(modelcheck)

reslm <- ggplot(bounce, aes(y = resid(lmfit), x = county, color = county)) +
  geom_violin() + theme_bw() + geom_jitter()+
  labs(title = "Residuals of the linear model by county",
       x = "county",
       y = "Residuals")

reshm <- ggplot(bounce, aes(y = resid(hmfit2), x = county, color = county)) +
  geom_violin() + theme_bw() + geom_jitter()+
  labs(title = "Residuals of the mixed models by county ",
       x = "county",
       y = "Residuals")
ggarrange(reslm, reshm, nrow = 1, ncol = 2)

check_model(hmfit2)

summary(hmfit2)
#sjt.lmer(hmfit2)
fixed_ci <- fixef(hmfit2)['(Intercept)'] + c(-2,2) * se.fixef(hmfit2)['(Intercept)']
fixef(hmfit2)
ran.ci <- tibble(County = rownames(ranef(hmfit2)$county),
                 lower = round((ranef(hmfit2)$county + -2 * se.ranef(hmfit2)$county),1),
                 upper = round((ranef(hmfit2)$county + 2 * se.ranef(hmfit2)$county),1))
ran.ci <- kable_styling(kable(ran.ci,
caption = " Interval estimates of random effects for various counties"),
latex_options = c("hold_position", "striped" ))

loc.ci <- tibble(Location = rownames(ranef(hmfit2)$location),
                 lower = round((ranef(hmfit2)$location + -2 * se.ranef(hmfit2)$location),1),
                 upper = round((ranef(hmfit2)$location + 2 * se.ranef(hmfit2)$location),1))

loc.ci <- kable_styling(kable(loc.ci,
caption = " Interval estimates of random effects for various locations"),
latex_options = c("hold_position", "striped" ))

```