678_final

Yingnan Lyu

2022-12-05

Abstract

Bicycle sharing is a very popular industry in the last few years. Bicycle sharing not only brings a lot of convenience to the daily life of urban residents but also the creation of many bicycle-sharing companies has injected a lot of vitality into the market. In order to run the company better and get the maximum profit, many companies want to forecast the rental volume based on some specific data from the past, which will help them to plan more rationally. Thus, here comes the problem: what kind of data do bike-sharing companies need to collect to help them forecast future rentals? To figure out this problem, I built a multilevel model with group level 'month'and 'weekday'. The result indicates that in different months and weekdays there are different pattern for people to rent bikes. This report can be divided into 4 main parts: Introduction, Method, Result, and Discussion.

Introduction

In order to see the pattern of renting, I choose the data from a bike-renting company to analyze. The service provider collects the entire US market based on a few factors. It contains various details of weather conditions, temperatures, windspeed, year, month, day, and whether it was a holiday or not when the bike was rented by the customers. Thus, I want to apply a multilevel model to this data to determine drivers that can make bike renting numbers go up. They want to understand the factors affecting the demand for these shared bikes in the American market. Specifically, the company wants to know:what variables are important in forecasting the demand for shared bikes, and how well do these variables explain the demand for bicycles based on various weather surveys and styles of people.

I will model the demand for bike sharing with the available independent variables. Managers will use it to understand how the needs of different functions are changing. They can manipulate business strategies accordingly to meet demand levels and customer expectations. In addition, the model will be a great way for management to understand the demand dynamics of new markets.

Methods

Data Preprocessing

I found this data from the kaggle website(https://www.kaggle.com/datasets/shrutipandit707/bikesharing).

I download the data from the website and add two columns(casual and registered) together to get a new column to create the appropriate data frame named df2. The new column delegates the total number of the renting bikes in a day. In this case, this data frame is suitable for me to apply a multilevel model to see which factor can influence the renting in different group level.

Here is the glossary of terms:

column names	explanation
instant	Instant when Bike was rented
dteday	Date when Bike was rented
season	Season when Bike was rented
yr	Year when Bike was rented
mnth	Month when Bike was rented
holiday	If Bike was rented on a Holiday or a Working Day
weekday	WeekDay when Bike was rented
workingday	If Bike was rented on a Holiday or a Working Day
weathersit	In Which Weather season Bike was Rented
temp	Temperature description when Bike was rented
atemp	Atemp when Bike was rented
hum	Humidity when Bike was rented
windspeed	Windspeed when Bike was rented
casual	If user who rented the bike is non-registered user
registered	If user who rented the bike is a registered user
cnt	Count of times when Bike was rented
total_number	total number of the renting that day

Exploratory Data Analysis

I’ve got a dataframe with 730 observations and 17 variables. I choose ‘total_number’ as the output and pick 5 factors as predictors. Then, I did following analysis in order to determine which predictor to fit the multilevel model.

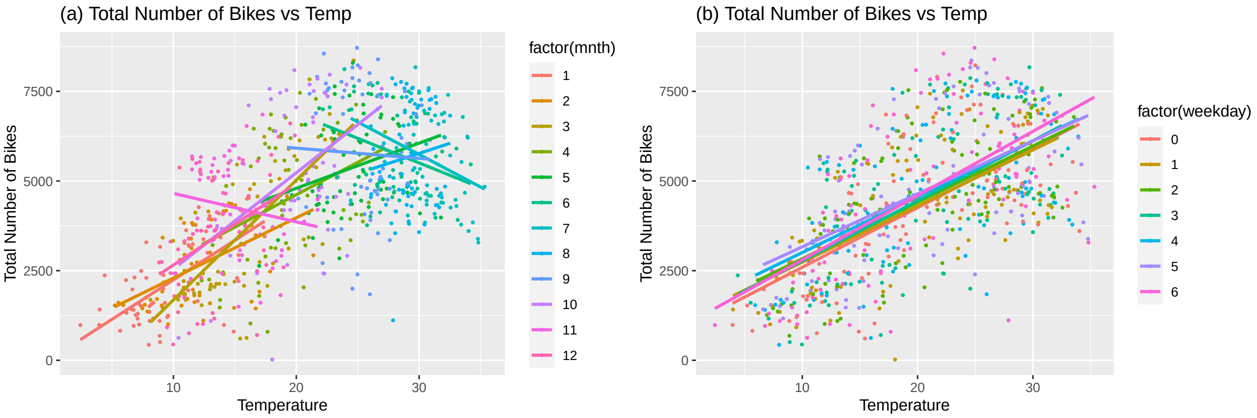


Figure 1: total_number vs temperature

The figure above shows the relationship between the total number of renting bikes and temperature in different group level. At the month level, some groups have the increasing trend but others have the decreasing trend. At the weekday level, the number of renting bikes will increase when the temperature goes up. In this case, we choose temperature as the random effect at the month level. In a similar way, the predictor atmep has the similar trends in this two levels. Thus, temperature and atemp can be the predictor at the month level but not at the weekday level.

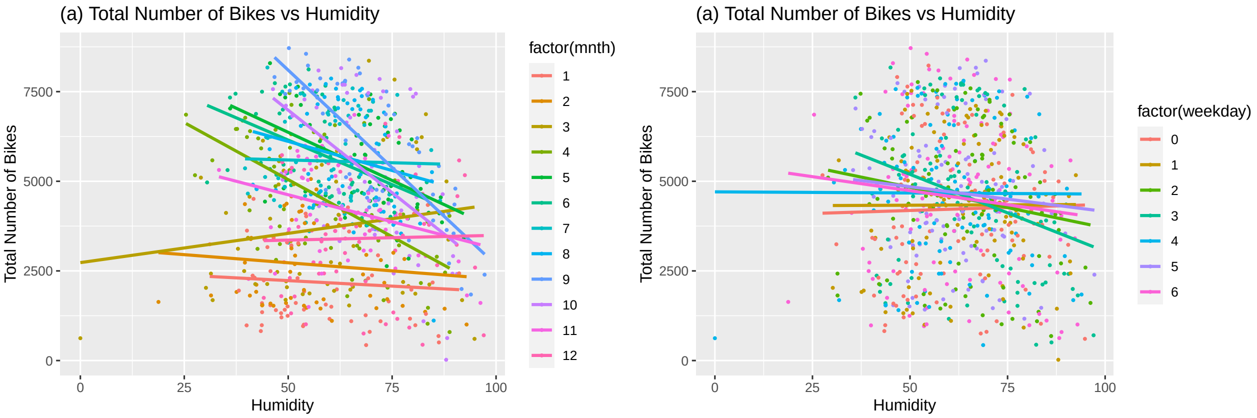


Figure 2: total_number vs humidity

This figure is about the correlation between the total number of bikes and huminity. There are two different kinds of trends at both the month level and weekday level. Thus, huminity can be included as the predictor at these two level.

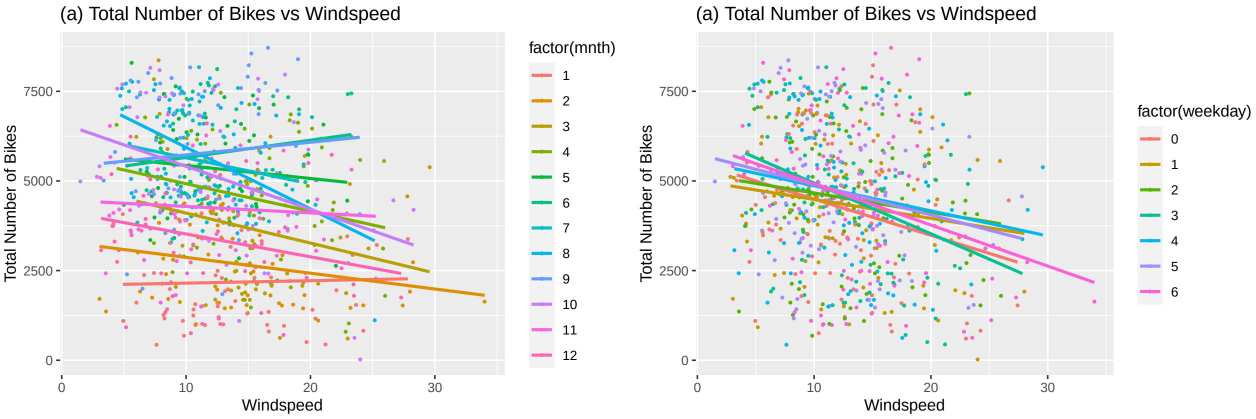


Figure 3: total_number vs windspeed

These two figures are about the relationsips between the total number of renting bikes and the windspeed at month level and weekday level. We can see from the figure(a), the trends differs from month to month, while the trends in different weekdays are similar. Therefore, windspeed can be a random effect at the month level to fit model.

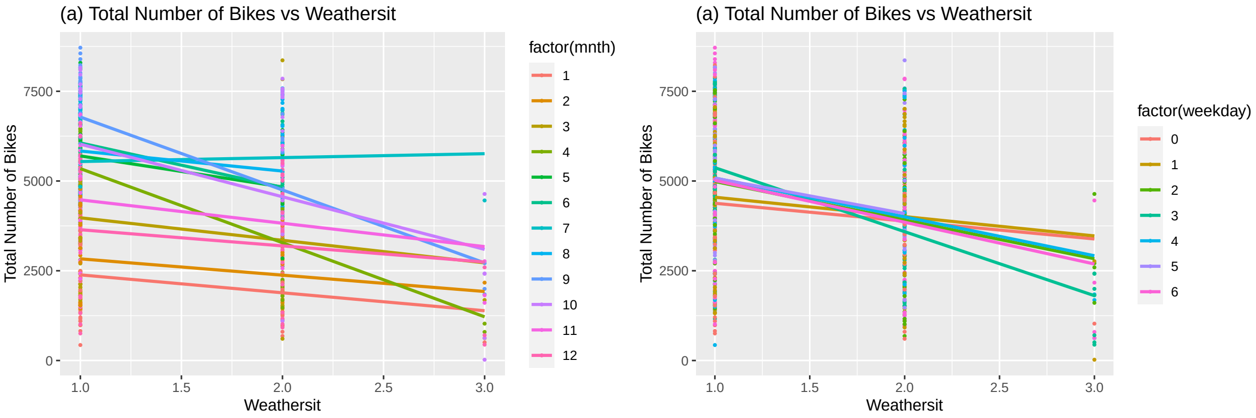


Figure 4: total_number vs weathersit

In figure(a), the mainly trend is decreasing but they are in varying degrees. In the figure(b), the trends are similar. In this case, I decided to use weathersit as a random effect at month level.

Now we have four predictors and roughly know their fixed effect. We're a few steps away from modeling.

Correlation of Data

Since different months and weekdays have a large impact on the model, I decided to use multilevel model to fit df2. In order to do the predictor selection, I did the Pearson correlation matrix.

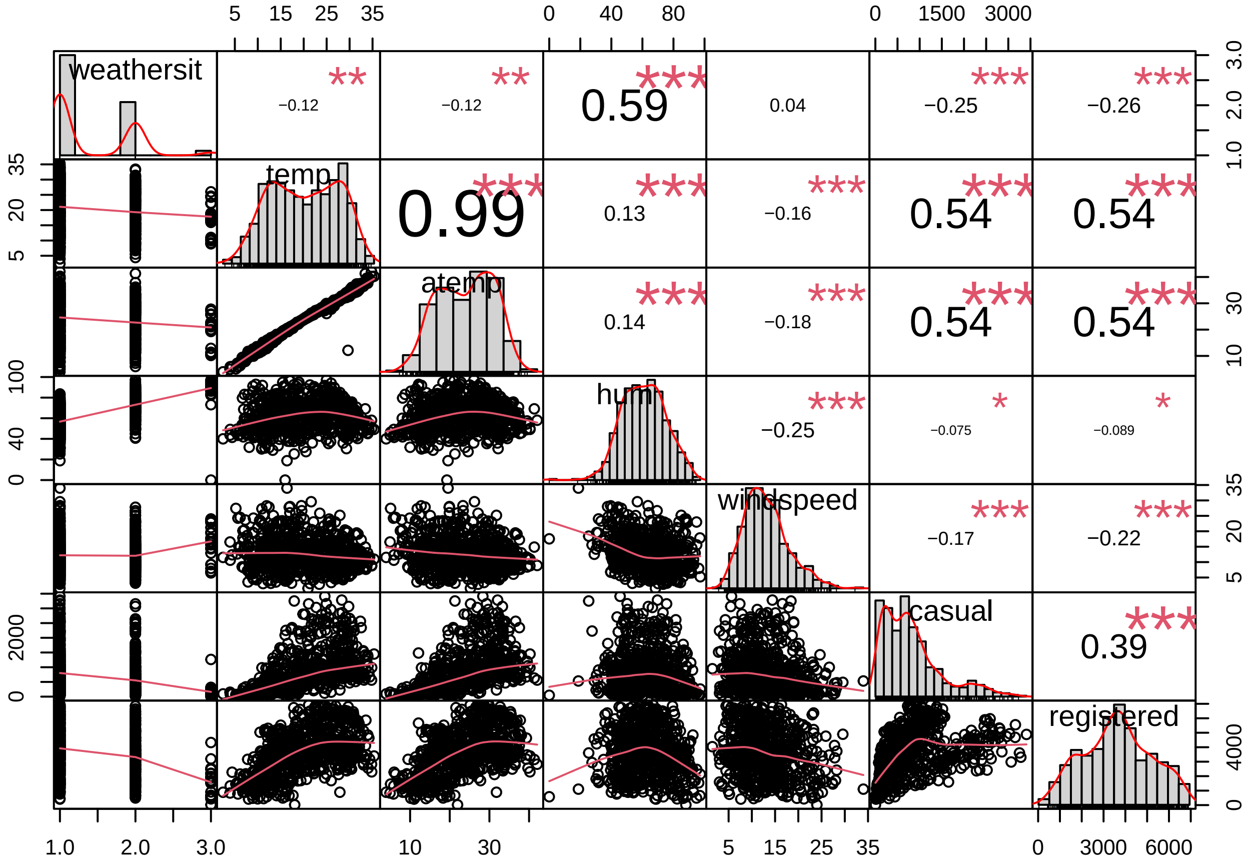


Figure 5: Correlation Matrix

This figure clearly shows the Pearson relationship between the variables. I choose .6 as the threshold to represent the variables which are highly related. Because the temperature and windspeed are the most important factors that can influent the renting number, I decided to keep them. From this matrix, .99 means temperature and atemp are highly related, which make me decide to drop one of them to fit the model. Therefore, we have weathersit, temperature, huminity and windspeed as ramdom effect to fit the model at month and weekday level.

Model fitting

Here is the function:

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	5877.552	885.032	17.282	6.641	3.83e-06 ***
temp	153.397	36.264	13.526	4.230	0.000903 ***
hum	-40.806	10.031	14.094	-4.068	0.001137 **
windspeed	-65.100	12.723	9.187	-5.117	0.000591 ***
weathersit	-265.059	107.490	675.340	-2.466	0.013915 *

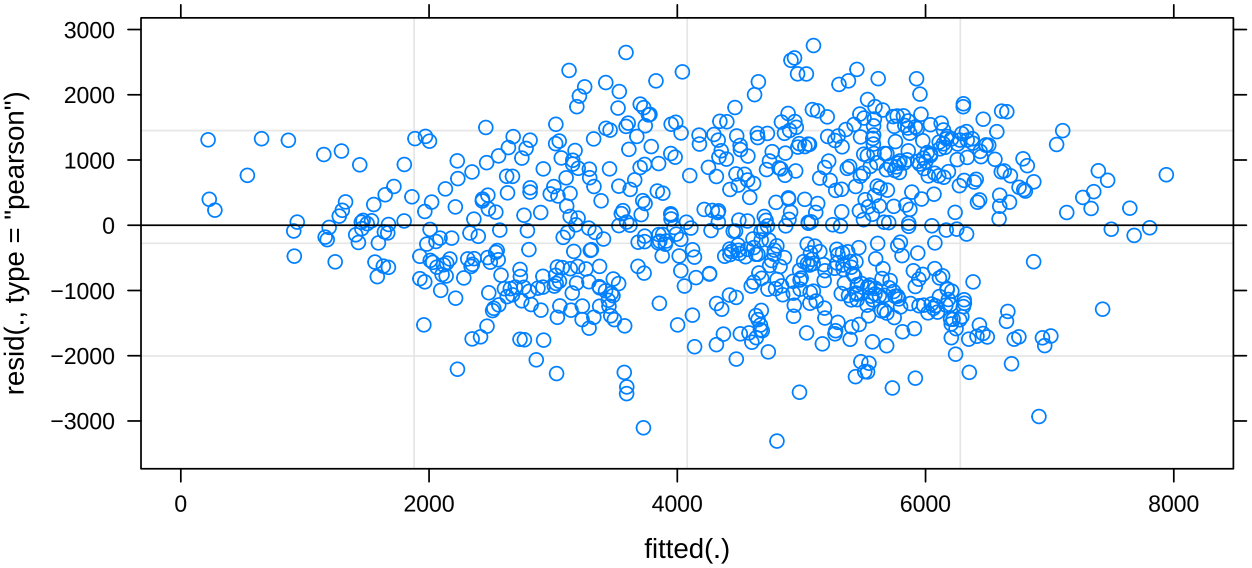
The chart above is the summary of the fixed effect and all the variables are considered as significant at .6 level. We can see it more clearly in the next figure.

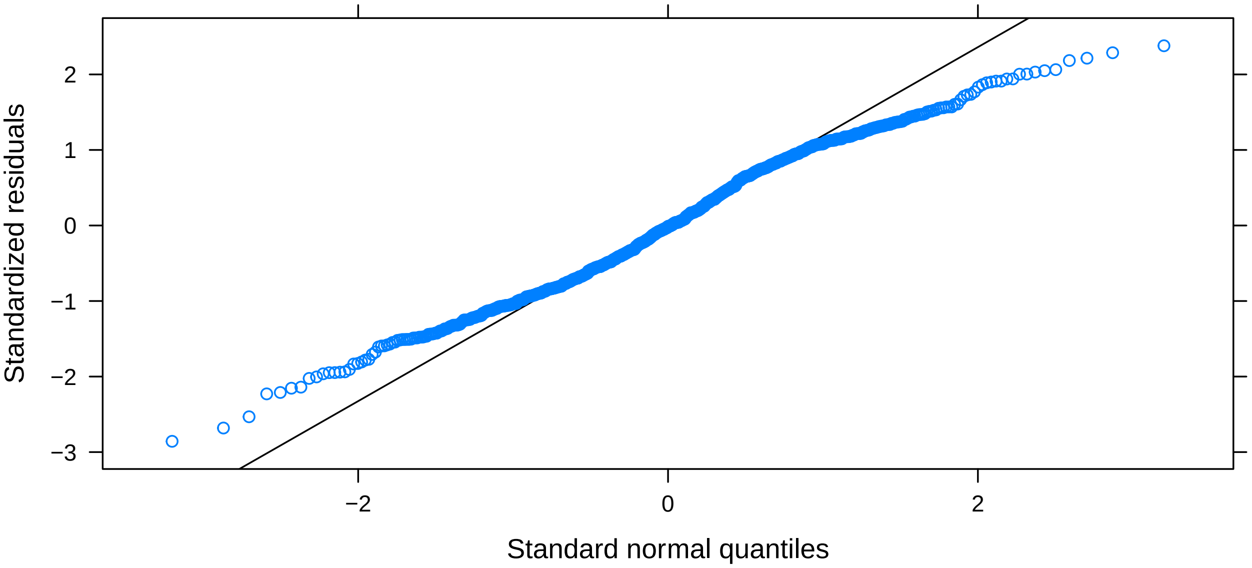
Besides, the following tables are summary of the ramdom effects. The first table is about random effect at the month level, and the second table is at the weekday level.

```
##    (Intercept)    temp    hum windspeed
## 1     -3099.78  108.68   9.85     -1.69
## 2     -3005.66   58.45  22.17    -27.25
## 3     -4738.47  151.75  27.55    -15.32
## 4      -165.29   35.01 -13.20    -16.05
## 5      1070.40  -34.52  -8.60      8.38
## 6      5564.93 -239.41  -4.81     39.10
## 7      3133.69 -219.94  31.41     14.50
## 8     -1342.81  -33.21  29.46    -45.65
## 9      3828.88    8.58 -62.46     42.47
## 10      708.37  103.30 -38.95     14.71
## 11      598.80  -34.51  -4.53      1.34
## 12    -2553.06   95.84  12.10    -14.52
```

```
##   (Intercept)   hum windspeed
## 0     -452.30  5.15      4.24
## 1     -153.83  1.75      1.44
## 2      109.67 -1.25     -1.00
## 3      361.17 -4.11     -3.39
## 4      -87.99  1.00      0.82
## 5      171.17 -1.95     -1.60
## 6       52.10 -0.59     -0.50
```

Model checking





According to the residual plot and residual Q-Q plot, the mean value of the residuals is approximately equals to zero. In the Q-Q plot, most of the points exept the tail ones are on the normal distribution line. Thus, the normality check is acceptable.

Result

Interpretation

This is the fomular of fixed effect:

$$log(total_number) = 5877.6 + 153.4 \times temp - 40.8 \times hum - 65.1 \times windspeed - 265.1 \times weathersit$$

Then add random effect of to the intercepts and slopes to get the model. Let's take July as an example.

$$log(total_number) = 9011.2 - 66.5 \times temp - 9.4 \times hum - 50.6 \times windspeed - 265.1 \times weathersit$$

All parameters in this model are negatively correlated with the dependent variable. This means that high temperatures, increased rainfall, and high wind speed will reduce the number of bike share rentals. In this model, the 730 observations are divided into two groups, one based on month and the other based on weekday. Four variables were selected from 17 variables to fit the model, namely temperature, humidity, wind speed, and weather type, and the EDA images were used to determine which variables were random effects.

According to this model, each degree increase in temperature decreases the number of rentals by 66.5, each unit increase in humidity decreases the number of rentals by 9.4, and each unit increase in wind speed increases the number of rentals by 50.4, with a difference of 265 between the different weather types.

Discussion

Through the construction of the model, we can clearly see the influence of various external factors on car rental volume. However, in this report, we mainly study the effect of weather on car rental volume, but it is known that not only weather affects car rental volume. Therefore, if bike-sharing companies want to better predict the number of rentals and develop business plans, they need to collect more data and analyze whether other types of factors have a significant impact on the number of rentals, such as epidemics, holidays, major events in the city, etc. The more variables and data available, the more profitable it will be to forecast car rentals, the more markets it will open up, and the more convenient it will be for citizens.

Appendix

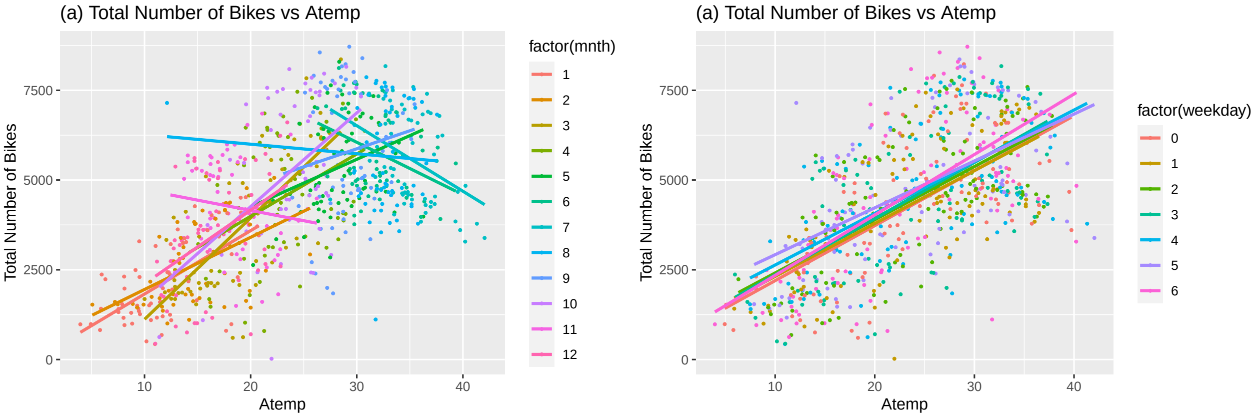
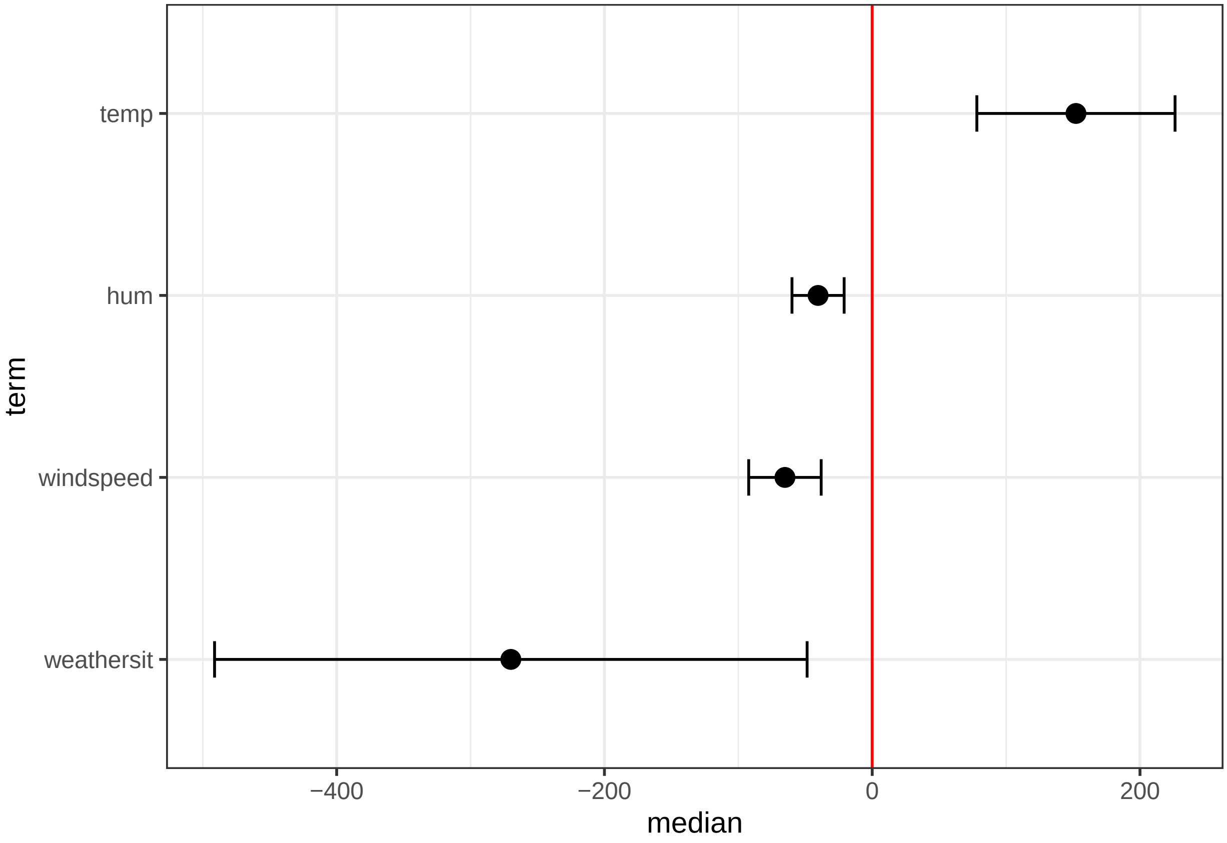


Figure 6: total_number vs atemp

```
plotFEsim(FEsim(multilevel_model, n.sims = 300), level = 0.95, stat = 'median', intercept = FALSE)
```


temp
hum
windspeed
weathersit
term
median