Prediction…

Sophie_Berube

9/28/2017

Introduction

Since the start of the nineteenth century when Legendre and Gauss published their least squares method, an early form of regression, as a way to describe bodies orbiting aroud the sun, statistical methods have been applied to a wide range of scientific problems. During the early 1800s through the beginning of the twentieth century, the notion of specialization in research and discovery as we understand it today did not exist. Francis Galton a british scientisit and father of the term "regression", first noticed the phenomenon of regression to the mean in a biological context. In this sense, statistics informed applied science as much as applied science informed statistical advances. However, in the latter half of the twentieth century with an explosion of tecnhology came increased scientific and mathematical specialization. People making siginificant advances in statistical methodology are now almost exclusively trained statisticians who can be brought on to applied scientific projects as co-investigators but do not conduct primary research in the natural or physical sciences the way those in the nineteenth century might have.

As a result of this fragmentalization of scientific research, methodological advances are made often in the statistical community and published in methodological papers. Oftentimes, software or computer code is subsequently developed to make the methodology easily applicable to a variety of scientific problems. After this process, the methodology begins to appear as a means of statistical analysis in various applied scientific fields ranging from neuroscience to genetics and genomics and even to physical sciences like astronomy.

The following analysis proposes a predictive model (imprecise or incorrect wording??) that attempts to find the next major statistical methodologies that will appear in applied science journals. More specifically, we (nobody else is writing this with me but we is more common in journals??) hypothesize that if the number of statistical methodology papers discussing a particular method increases, then this method after some lag time, will appear with greater frequency in applied scientific articles as a method of data analysis.

Methods and Materials

We will begin by defining a statistical methodology paper as a paper where the first or last author, are affiliated with a biostatistics, statistics or mathematics department? (posibly will change to one published in plos computational biology). Need to think about defining what is an increase in frequency of methodology in methodological papers, could use IQR? or something else?

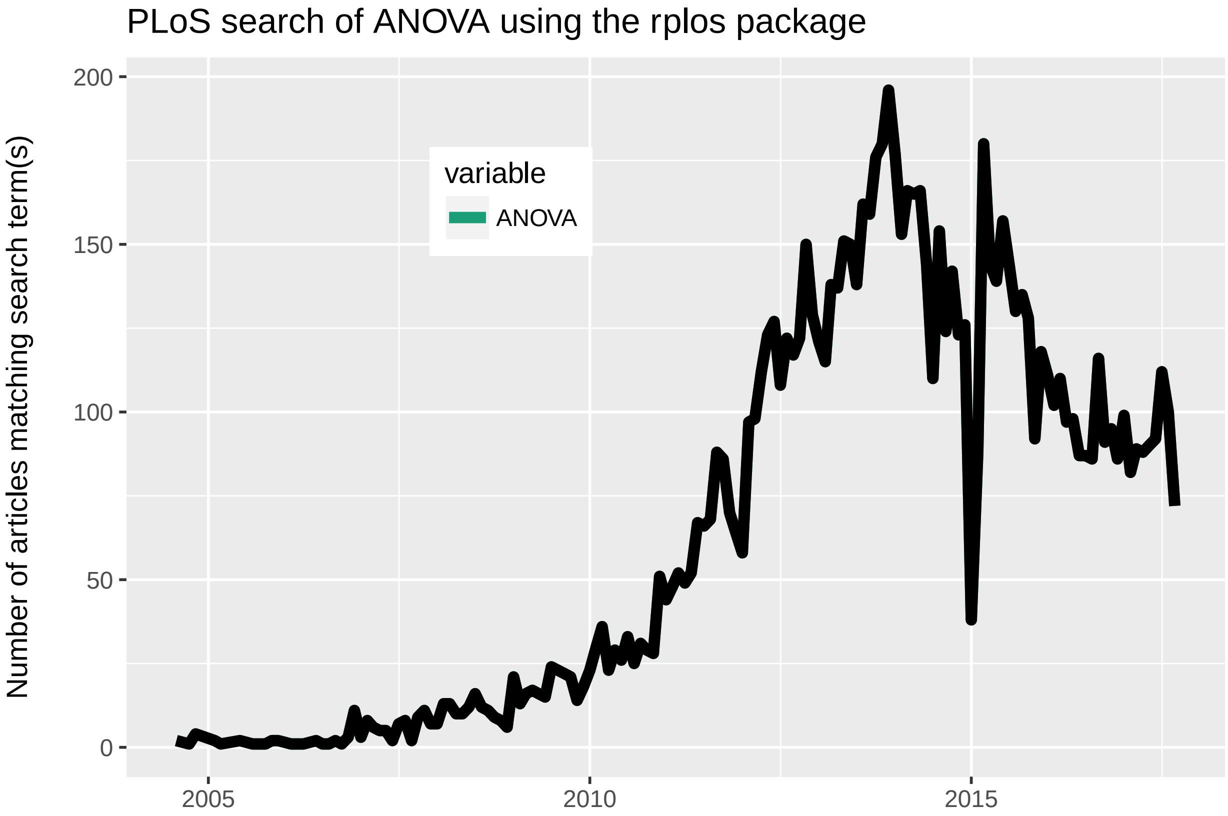
Need some ROC curve or positive predictive value for how well the mention of a statistical method in the materials and methods section will predict the use of the actual method in the applied paper? How will this error be represented, confidence intervals??

False Discovery Calculation

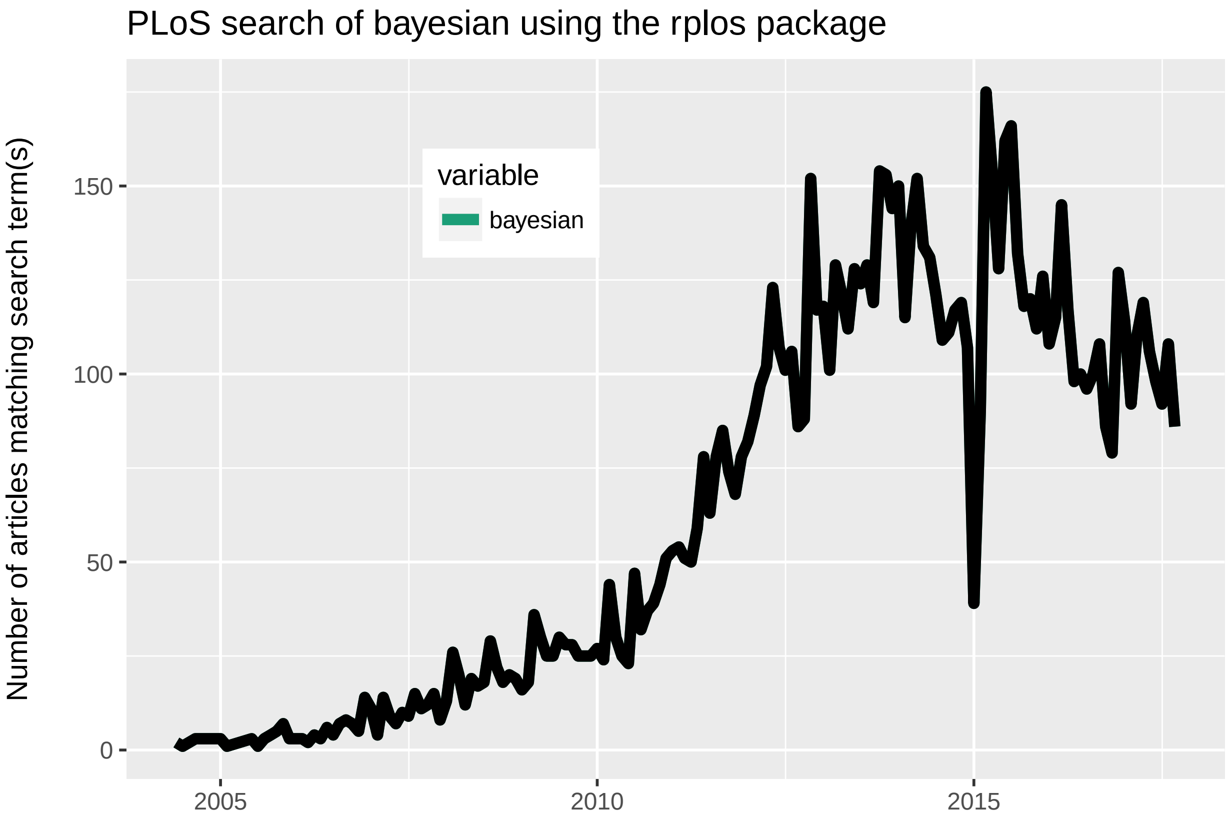
In order to assess how well the presence of certain key words in the body of an article correlate with the actual use of a particular statistical methodology in the paper a false discovery rate was computed using the `rplos` function `highplos`. More specifically, for each of the 10 key words outlined above, a sample of 10 articles that mentioned the word were found using the `highplos` function. Then the article was visually inspected to see whether the statistical methodology of interest was actually used in the study. Out of 100 articles observed, 17 contained a key word without using that particular techinque in the study. Thus the false discovery rate is roughly 17%.

Very preliminary result plots (maybe bleeding over into EDA?)

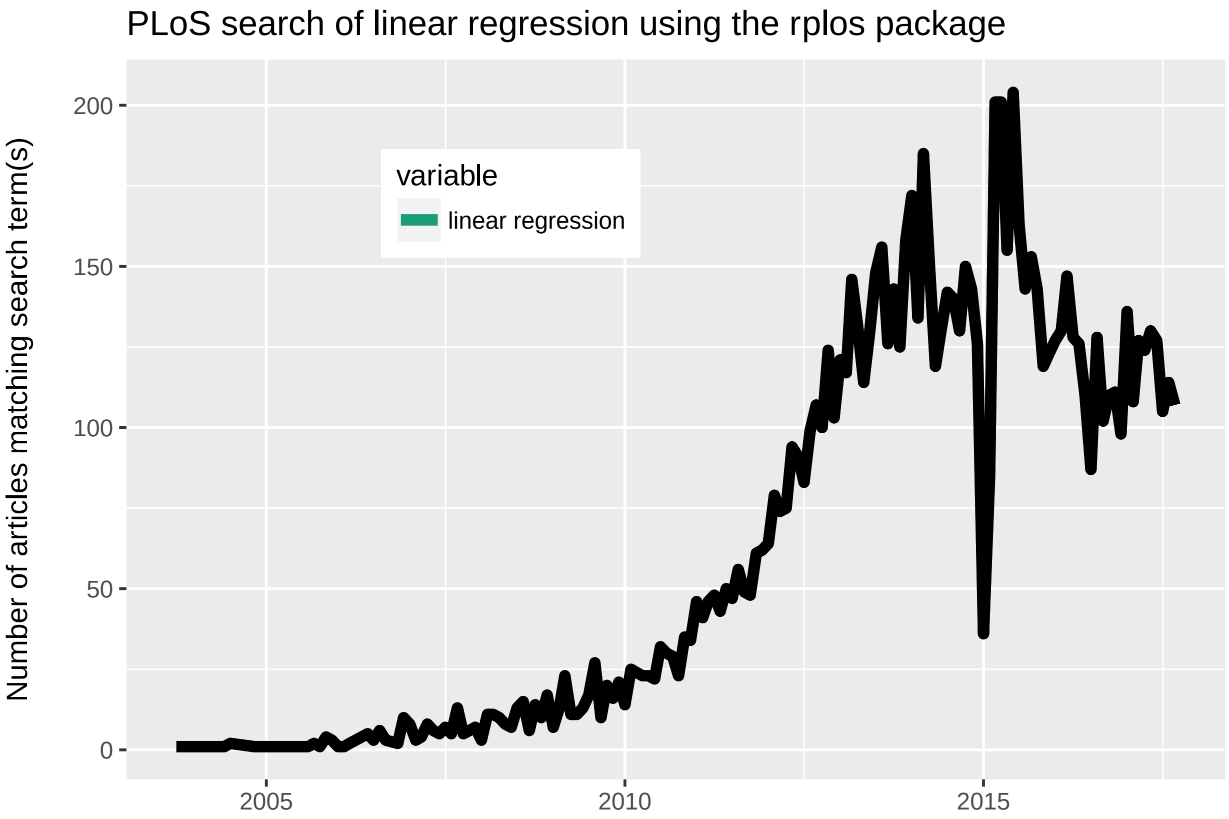
```
library(rplos)
library(ggplot2)
ANOVA_time<- plot_throughtime(terms="ANOVA",limit=10000) + geom_line(size=2,color='black')
ANOVA_time
```



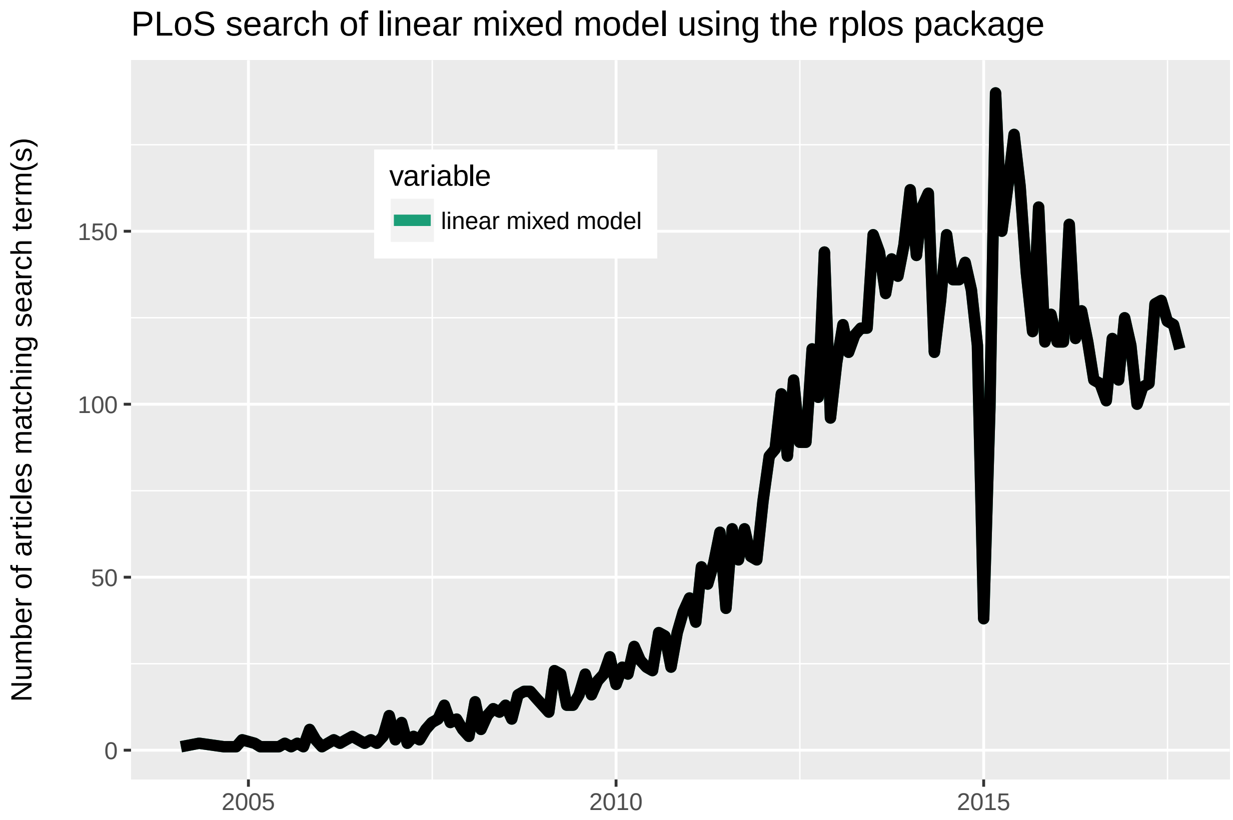
```
bayes_time<- plot_throughtime(terms="bayesian",limit=10000) + geom_line(size=2,color='black')
bayes_time
```


PLoS search of bayesian using the rplos package

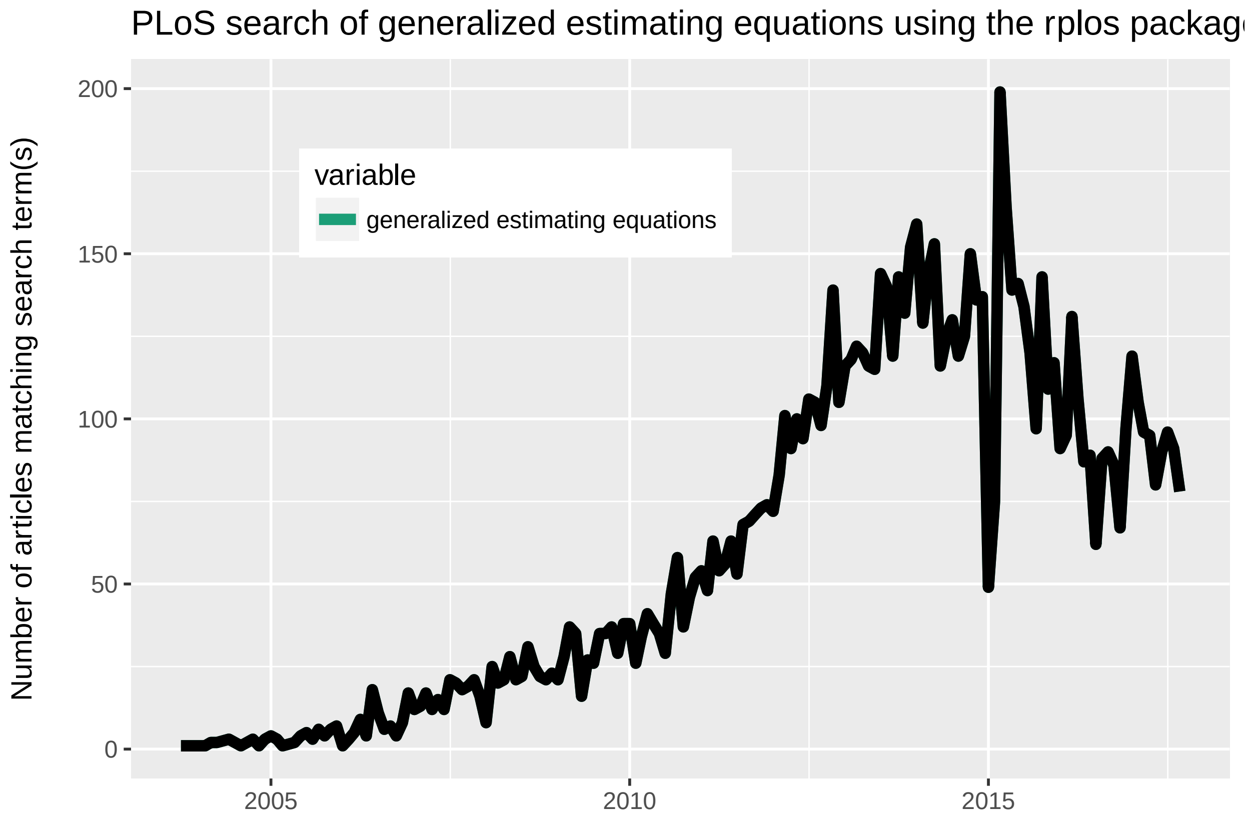
```
linearregression_time<- plot_throughtime(terms="linear regression",limit=10000) + geom_line(size=2,colo
linearregression_time
```


PLoS search of linear regression using the rplos package

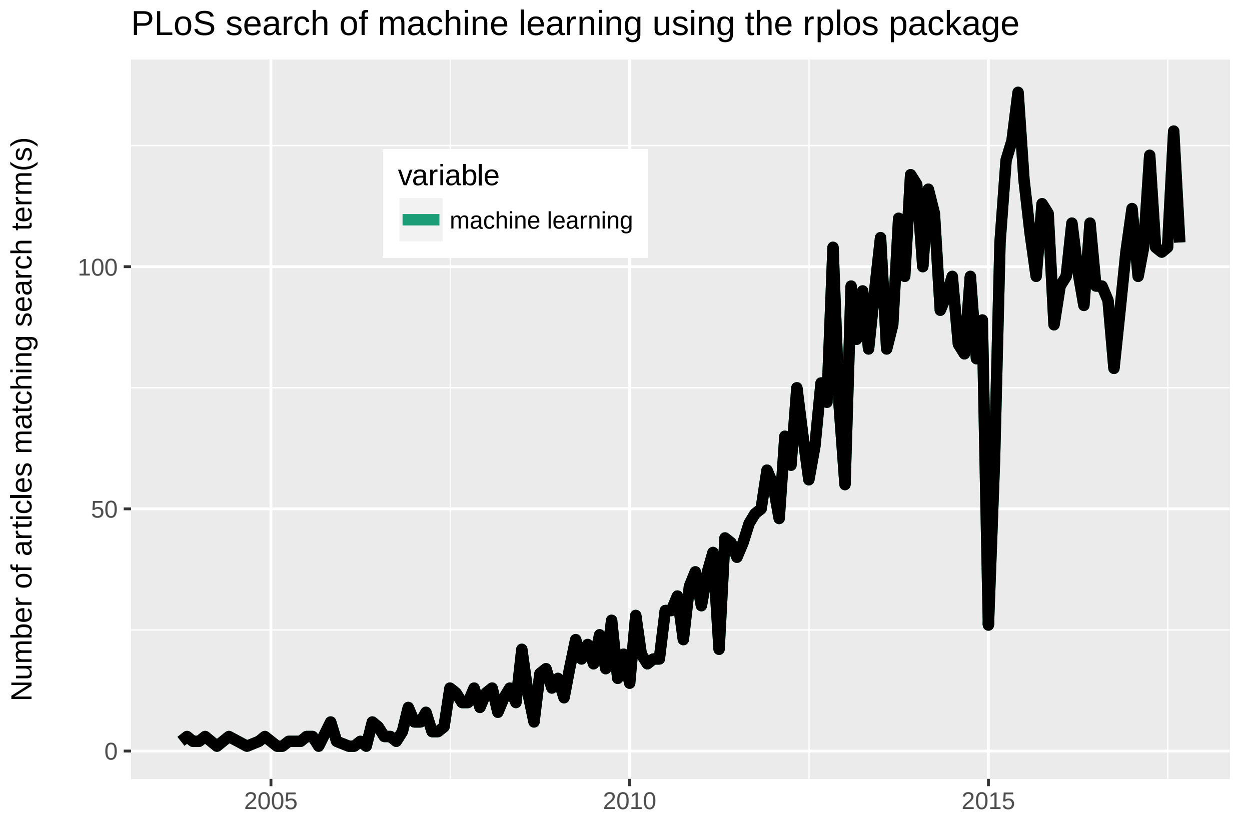
```
lmm_time<- plot_throughtime(terms="linear mixed model",limit=10000) + geom_line(size=2,color='black')
lmm_time
```



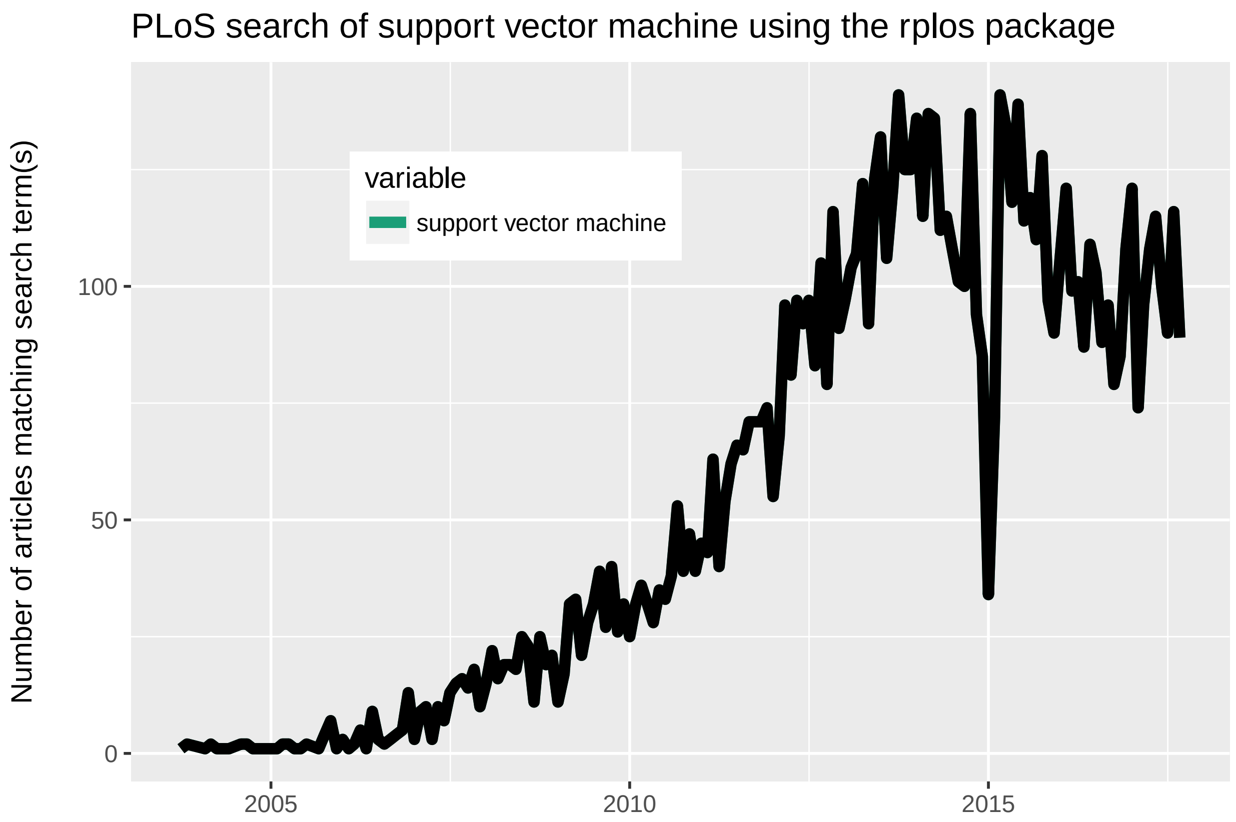
```
gee_time<- plot_throughtime(terms="generalized estimating equations",limit=10000) + geom_line(size=2,col
gee_time
```



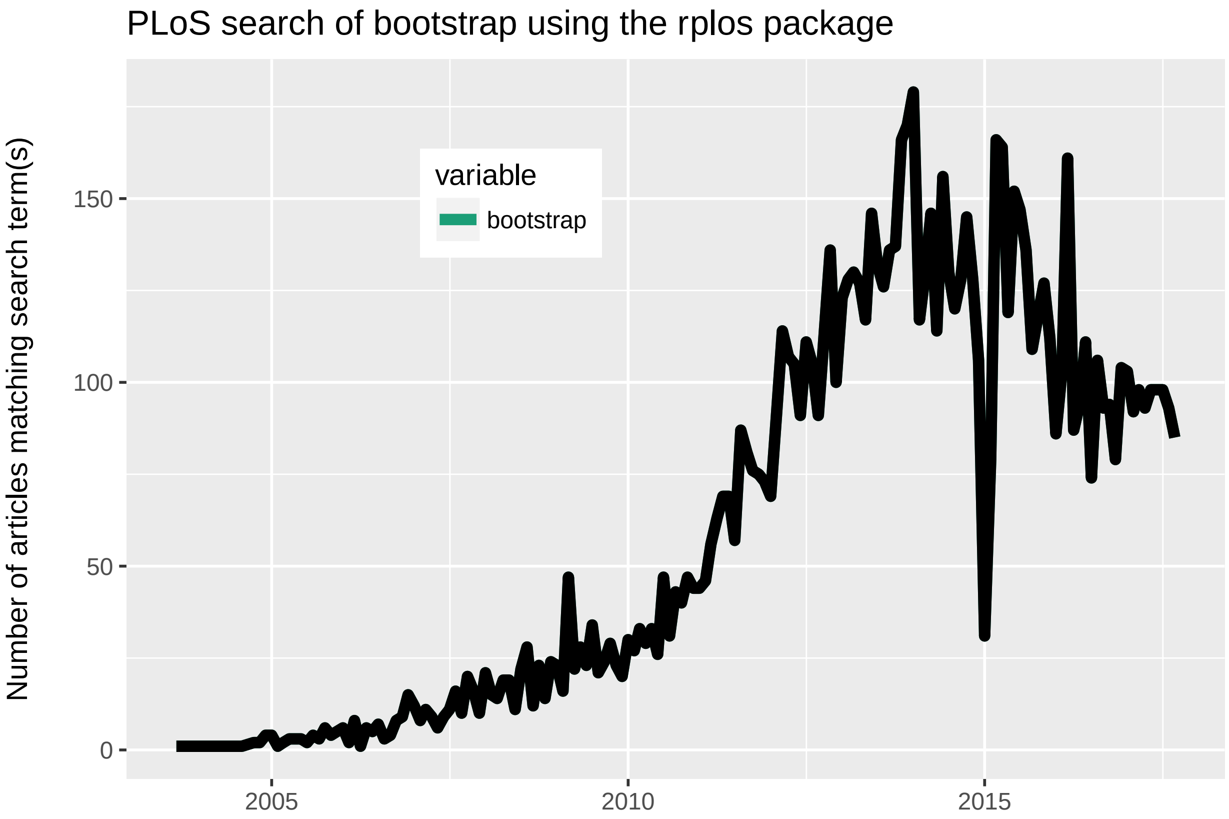
```
machinelearning_time<- plot_throughtime(terms="machine learning",limit=10000) + geom_line(size=2,color=
machinelearning_time
```



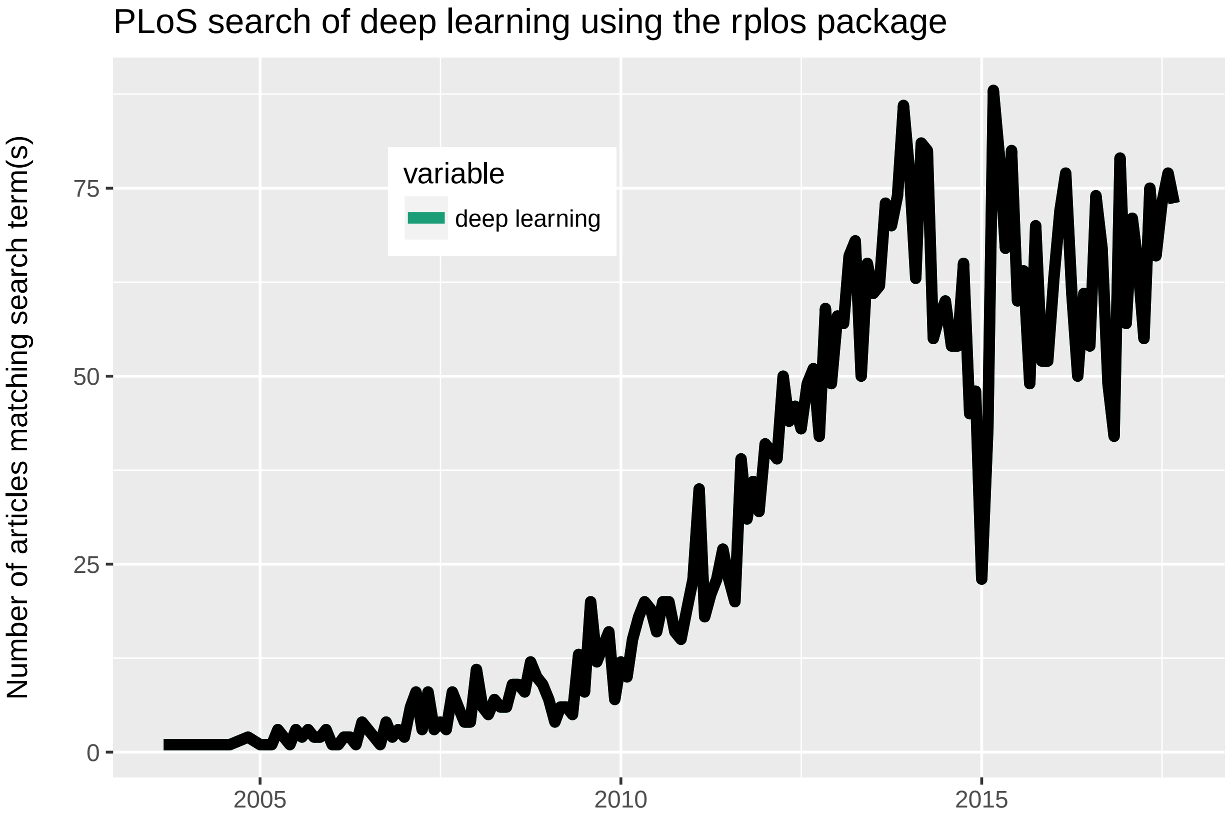
```
svm_time<- plot_throughtime(terms="support vector machine",limit=10000) + geom_line(size=2,color='black
svm_time
```



```
bootstrap_time<- plot_throughtime(terms="bootstrap",limit=10000) + geom_line(size=2,color='black')
bootstrap_time
```



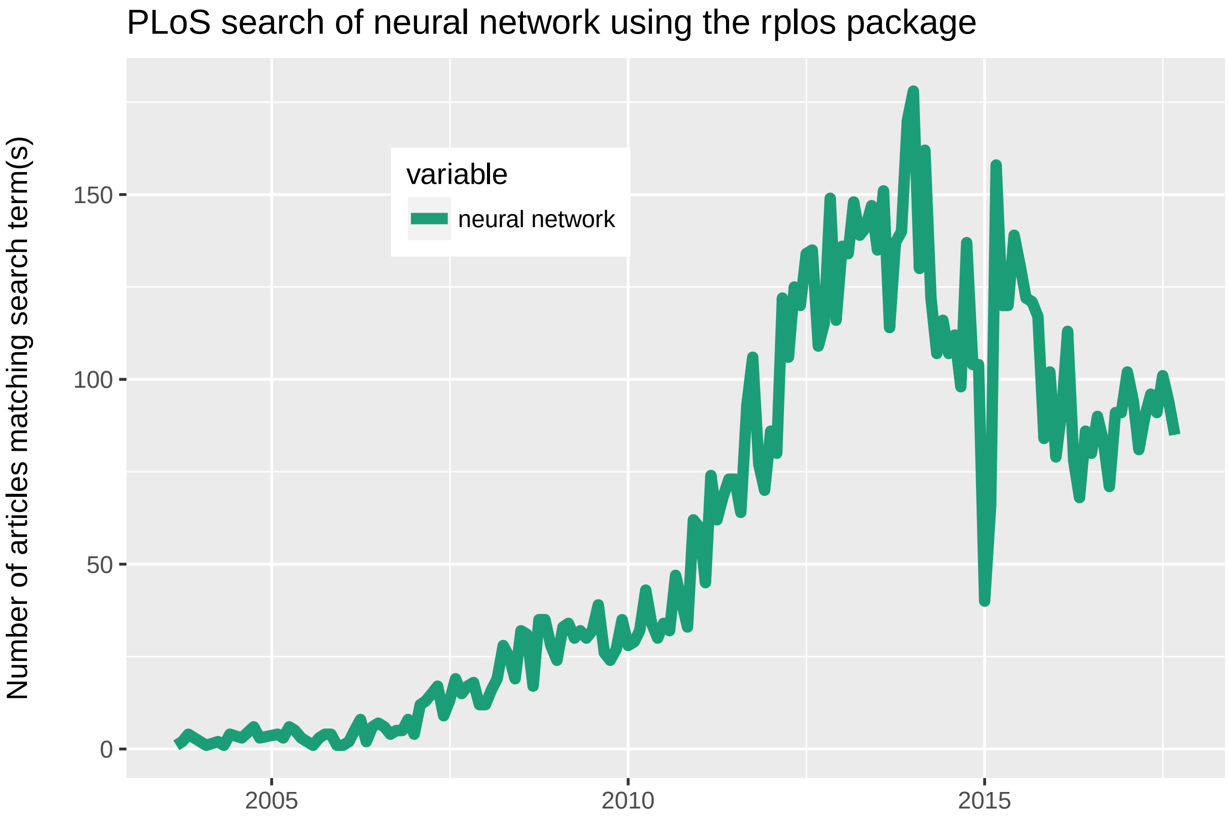
```
deeplearning_time<- plot_throughtime(terms="deep learning",limit=10000) + geom_line(size=2,color='black
deeplearning_time
```



```
neuralNet_time<- plot_throughtime(terms="neural network", limit=10000)
geom_line(size=2,color='black')
```

```
## geom_line: na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

```
neuralNet_time
```


PLoS search of neural network using the rplos package
Number of articles matching search term(s)
variable
🟩 neural network