Data Analysis Final Exam

Andrew Resnikoff
aresniko
36-401 Modern Regression

Abstract

Since blogging has become popular, it is interesting to see how the number of comments vary for different blog posts so that we can understand what drives web traffic to certain posts rather than others. In this investigation, multivariate linear regression was used to determine what affects the number of comments a blog post gets 24 hours after a given basetime. Starting with a sample of 668 blog posts and 27 potential predictors, we eliminated outliers and insignificant variables to build a final model. In this model, we discovered that certain words can either cause an expected increase or decrease in the number of comments in those 24 hours. We also found that posts with more comments 24 hours before the basetime (for the source average or the post itself) are expected to have more comments after the basetime.

Introduction

Ever since the proliferation of the internet, individuals now have the power to share whatever they want to whoever is curious enough to pay attention. People use blogs to share information, but currently it is difficult to predict exactly which posts receive the most traffic. In this analysis, we will explore different qualities of blog posts and determine what affects the number of comments a blog post gets 24 hours after a given basetime. We expect that the source of the post will have more of an effect on the number of comments than the links and that there may be an interaction between basetime day and the number of comments 24 hours before the baseline. We also believe that longer posts will have fewer comments, blogs posted on Sunday or Monday will have more comments and blogs will an informative word will have more comments.

Exploratory Data Analysis

We are given a dataset with the following variables: the number of comments in the next 24 hours after the base time (response), number of comments in the first 24 hours after posting, 24 hours before the bastille and the total number of comments before the bastille for this blog post, (the average of these values) for the blog post source and the links in the blog. We are also given the length of the blog post, the age of the blog post, the day of the week the blog was posted and the day of the baseline, whether or not words 1-10 appear in the post. Lastly, we are given the number of parent pages for the blog post, and the min,max and average number of comments these parent pages received. In total, there are 668 blog post observations in the dataset. Tables 1, 2 and 4 below show summary statistics for all of the variables we are given. Also, in Figures 1 and 2 below the tables, we have histograms for continuous variables and bar graphs for categorical variables respectively. Notice that in the histograms, other than for *post.age*, all of the variables are skewed right. This suggests we may later find outliers in the data for observations that have a lot of comments in the next 24 hours. From the bar graphs and summary statistics, we see that most of the words do not appear in many blog posts (if they even appear at all). The day of posting and basetime each seem to be pretty evenly spread.

Table 1 Key

variable name	comments.next24	comments.prebasetime	comments.prev24	comments.first24
key in table	a	b	c	d
variable name	source.avg.prebasetime	source.avg.prev24	source.avg.first24	links.prebasetime
key in table	e	f	g	h
variable name	links.prev24	links.first24	length	post.age
key in table	i	j	k	l
variable name	n.parents	parents.avg	parents.min	parents.max
key in table	m	n	o	p

Table 1: Summary Statistics for Continuous Variables

	a	b	c	d	e	f	g	h
Min	0	0	0	0	0	0	0	0
Max	1370	1594	1970	1262	546.6299	231.5906	442.5171	15
Median	0	3	1	3	10.6307	4.0849	9.7769	0
Mean	12.5569	47.9027	20.1796	42.5150	47.6083	18.6994	41.6596	0.5569
SD	79.2999	134.3224	71.3920	110.8584	99.4068	41.1247	82.7656	1.5789

	i	j	k	l	m	n	o	p
Min	0	0	0	0	0	0	0	0
Max	10	15	37236	72	6	83	0	83
Median	0	0	1635	35	0	0	0	0
Mean	0.2470	0.506	2544.898	34.4177	0.0928	0.2829	0	0.3368
SD	0.9828	1.4971	3475.494	20.7373	0.5369	4.0583	0	4.2608

Table 2: Summary Statistics for Yes/No categorical variables

	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
Yes	1.9461%	0%	0%	4.491%	3.2934%	0%	2.6946%	7.6347%	0%	2.6946%
No	98.0539%	100%	100%	95.509%	96.7066%	100%	97.3054%	92.3653	100%	97.3054%

Table 3: Summary Statistics for (multiple category) categorical variables

	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
basetime.day	14.6707%	10.479%	12.2755%	13.7725%	15.4192%	16.018%	17.3653%
post.day	8.6826%	15.7186%	16.019%	16.9162%	15.7186%	14.6707%	12.2754%

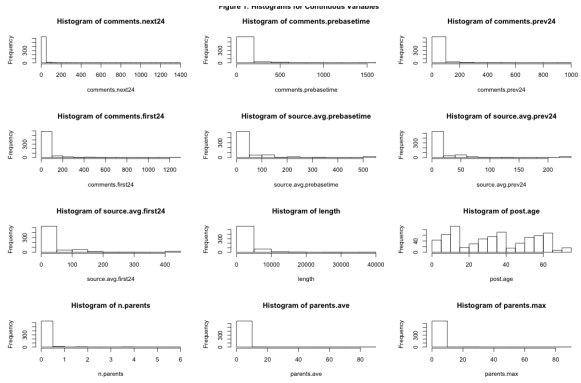

Figure 1: Figure 1

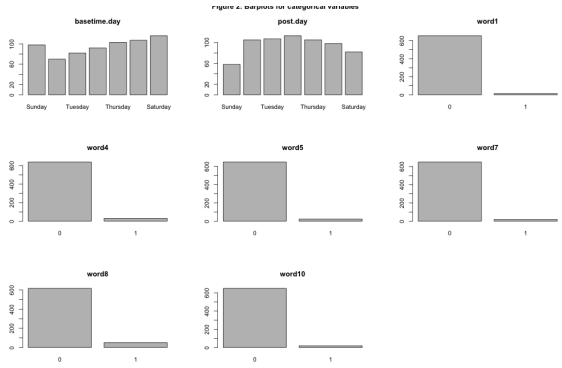

Figure 2: Figure 2

Looking at the pairs plot in Figure 3, we see that there are many significant relationships between variables. Specifically, *comments.next*24 has significant positive linear relationships with *comments.prebasetime*, *comments.prev*24, *comments.first*24, *source.avg.prebasetime*, *source.avg.prev*24, *source.avg.first*24, *links.prebasetime*, *links.prev*24, *links.first*24 and *length* (correlations of .19, .28, .20, .52, .54, .51, .16, .23, .17 and .12 respectively) as well as a significant negative linear relationship with *post.age* (correlation -.18). There are also many significant linear relationships among the predictor variables. Some notable relationships are *comments.prebasetime* and *comments.first*24 (correlation .99), *source.avg.prebasetime* and *source.avg.prev*24 (correlation 1.0), *source.avg.prebasetime* and *source.avg.first*24 (correlation 1.0), *links.prebasetime* and *links.first*24 (correlation .99) and lastly *parents.ave* and *parents.max* (correlation .99). These relationships, as well as all other relationships between variables will be explored further during multicollinearity and interaction testing.

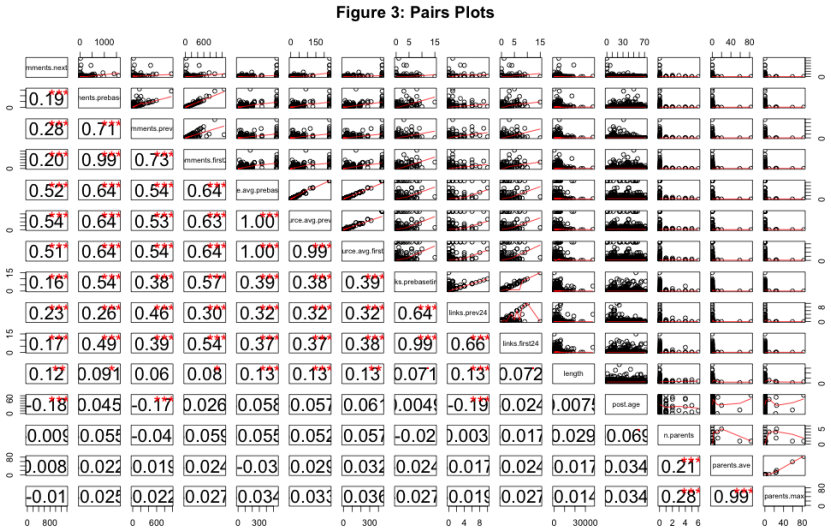

Figure 3

Model Selection

The model selection process starts by initializing a basic model, with *comments.next*24 as the response, and the remaining variables as predictors.

Next, we can remove predictors that offer no insight into the value of *comments.next*24 from our initial model. These predictors (which each have value 0 for all observations) are: *parents.min*, *word*2, *word*3, *word*6 and *word*9.

Now that those variables have been removed the model, we must appropriate code our categorical variables into the model. The "word" variables have already been coded as 0 for No and 1 for Yes (No is the reference category). This should give us valuable insight as to the effects of these variables as is. To test the hypothesis that blog posts that were posted on Sunday or Monday have more comments than blog posts on other days, we will code *post.day* as a new variable, *Post.day*, that has value 1 if the *post.day* is Sunday or Monday and 0 otherwise. For the variable *basetime.day*, a few potential coding methods were used. Variables were left as is, coded as weekend (Friday, Saturday, Sunday) or weekday (Monday, Tuesday, Wednesday, Thursday) and coded as groups based off of the boxplots in Figure 4. After trying all of these methods, while neither variable was ever a significant predictor for *comments.next*24, the weekend vs. weekday grouping had the lowest p-value, so we will use that grouping. Blog posts from the weekend will be coded as 1 and posts from the weekday will be coded as 0.

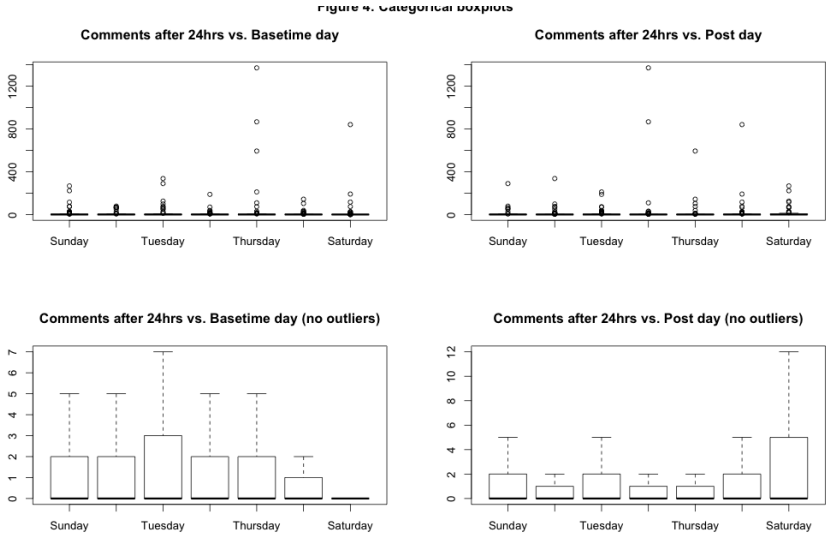

Figure 4

The high correlations between predictors as mentioned in **Exploratory Data Analysis** allows us to remove more unnecessary variables from our model. When variables have a correlation of 1.0 or .99, it means that one variable is a scalar multiple of another and it would be redundant to include both in the model. Notice that from the pairs plot in Figure 3, for the "comments", "source.avg" and "links" variables, the "prebasetime" has 1.0 or .99 correlation with "first24". Thus, we will select the "prebasetime" variables to include because these variables include the information in the "first24" models (and hence we have now excluded $comments.first24$, $source.avg.first24$ and $links.first24$ from the model). Additionally, $source.avg.prebasetime$ has a 1.0 correlation with $source.avg.prev24$ so we will also remove $source.avg.prev24$ from the model. To verify the validity of these procedures, we can perform partial f-tests.

We investigated the suggested interaction between $comments.prev24$ and $Basetime.day$ using the graph in Figure 5 and a chi-squared test for independence although neither had strong evidence for the inclusion of this interaction term. Additionally, we performed chi-squared tests on all pairs of variables and found many more significant interactions. After performing more partial F-tests, the significant predictors were found to be: $comments.first24$ and $links.prebasetime$, $links.prebasetime$ and $links.prev24$, and $links.prebasetime$ and $word8$.

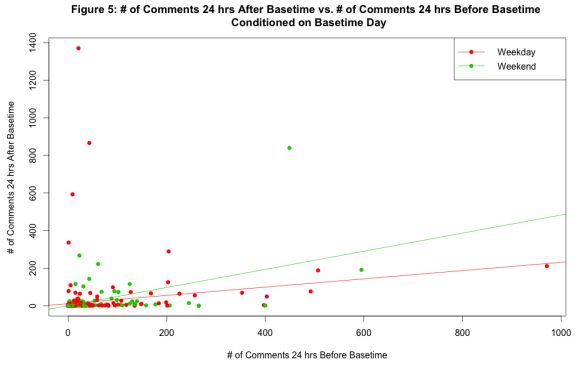

Figure 5

Diagnostics

Due to variable skews in the EDA, we suspected that outliers were present in our model. Specifically, it appeared that these outliers occurred for high values of *comments.next*24. After modeling the data, we tested for the presence of outliers using the deleted standardized residual technique. This method resulted in the removal of 9 observations (deleted standard residual values: -5.65, -4.88, -4.65, -4.46, 4.08, 4.39, 8.32, 8.37 and 23.46; t-score = 3.99). Many of these removed observations come from observations with many comments in the next 24 hours after the basetime, confirming our suspicions. Outliers found thorough other methods were not removed from the model.

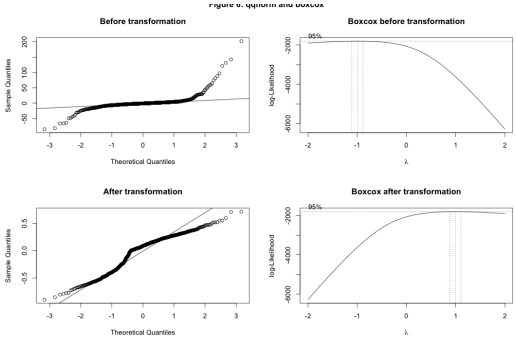

Figure 6

In Figure 6, we show the residual diagnostic plots being used to ensure that the assumptions for our model are met. The first qqnorm plot, testing for normality, indicates that our model does not currently meet the normality assumption and that a boxcox transformation is necessary. After shifting the response to make it positive, the initial boxcox transformation returns $\lambda = -1$. After performing the transformation, the qqnorm plot appears to have been worsened even though 1 is now included in the 95% confidence interval. When we compare the residual plots and independence tests in Figure 7, it is evident that the assumptions are better met with the pre-transformed model. In the pre-transform model, the residual plot seems like it could have mean zero and constant variance and there are no patterns in the residuals (as opposed to in the transform model where there is definitely a pattern in the residuals and thus is not randomly scattered). The residual plots for predictor variables (shown in Figures 8 and 9) are not too concerning and definitely allow us to continue on with the model.

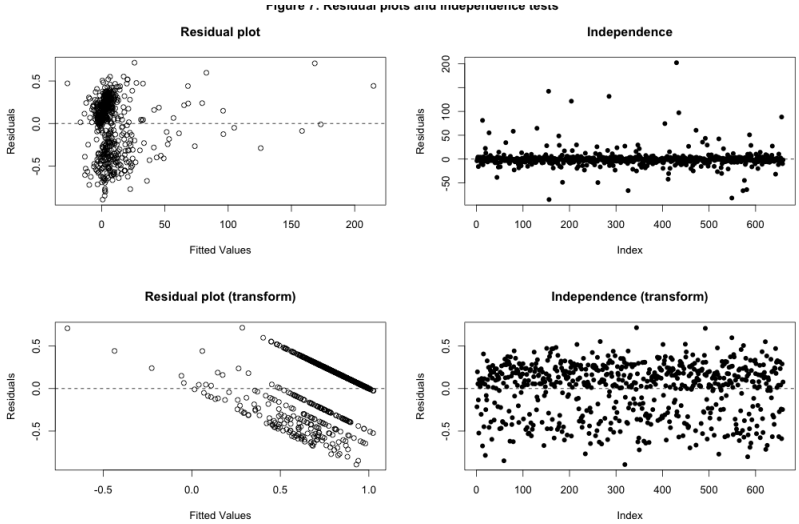

Figure 7

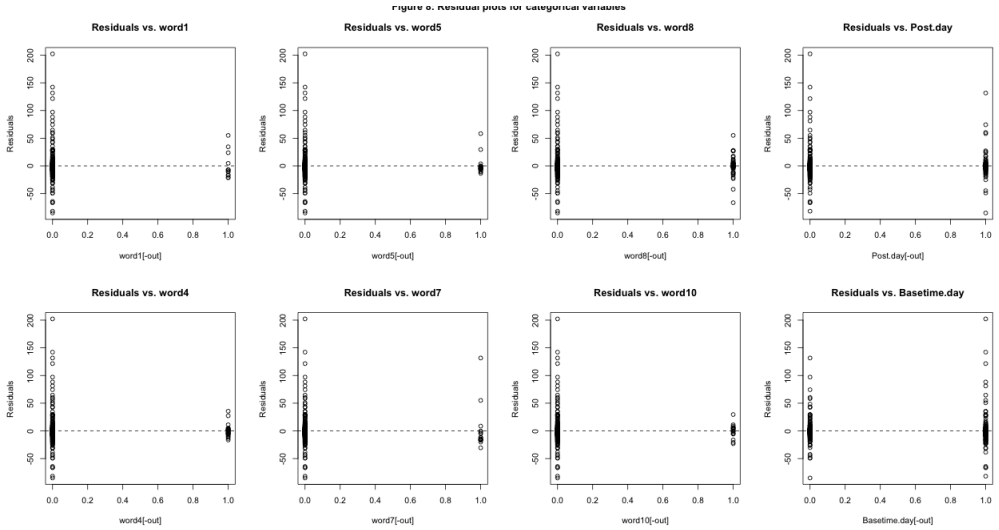

Figure 8

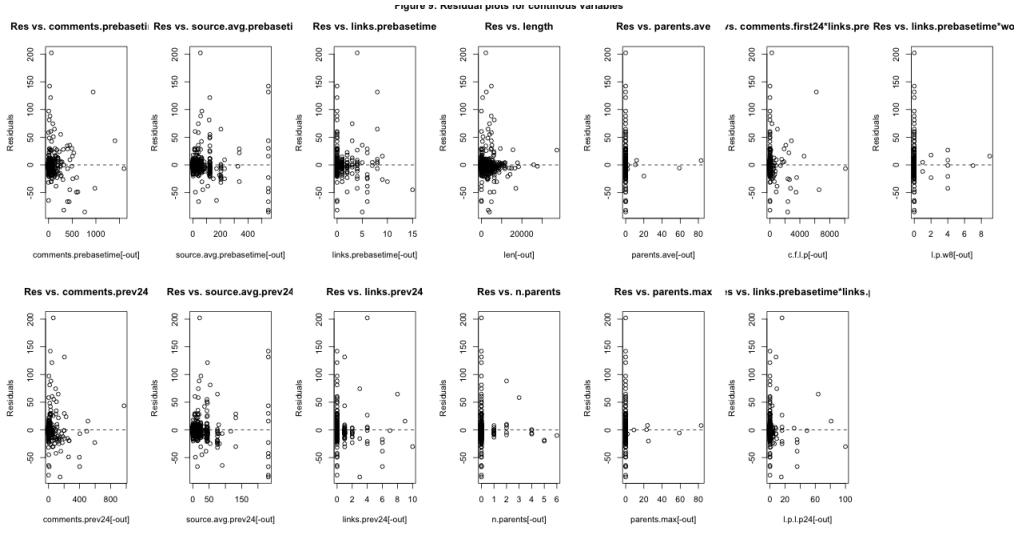

Figure 9

Model Inference and Results

Now that we have a model that meets the correct assumptions, we begin the model selection process. We can use a backwards step regression on our current model to help choose our model. Using the result of the backwards step model, we can choose which variables to try to eliminate using a partial F-test. After performing this F-test and failing to reject at the .05 level, we were able to remove 13 predictors from the model ($n.parents$,$links.prebasetime * links.prev24$, $parents.max$, $parents.ave$, $word4$, $word5$, $word8$, $word10$, $Basetime.day$, $Post.day$, $length$, $links.prev24$ and $links.prebasetime$). A table of the summary for our final model is shown below in Table 4.

Table 4: Final Model Summary

Variable	β (Std. Error)	t-score	P-value
(Intercept)	4.030528 (1.536662)	2.623	0.008923
comments.prebasetime	-0.109461 (0.014372)	-7.616	9.25e-14
*comments.prev*24	0.170395 (0.016729)	10.185	$<$ 2e-16
source.avg.prebasetime	-0.474290 (0.108997)	-4.351	1.57e-05
*source.avg.prev*24	1.631207 (0.268304)	6.080	2.05e-09
post.age	0.159062 (0.030531)	5.210	3.71e-07
*word*1	-0.118344 (0.037267)	-3.176	0.001566
*word*7	13.203045 (4.538878)	2.909	0.003751
$comments.first24 * links.prebasetime$	0.014222 (0.002023)	7.031	5.20e-12
$links.prebasetime * word8$	-5.123209 (1.390727)	-3.684	0.000249

In our final model, *comments.prebasetime*, *comments.prev*24, *source.avg.prebasetime*, *source.avg.prev*24, *post.age*, *word*1, *word*7, $comments.first24{*}links.prebasetime$ and $links.prebasetime{*}word8$ were significant predictors for *comments.next*24, the number of comments a blog post would receive 24 hours after the basetime. This model had $R_a^2 = .5074$. Notice that having more comments before the baseline (for the source average or the post itself) is associated with more comments (despite pre-basetime coefficients being negative, as increased prev24 will also increase prebasetime and prev24 has a greater coefficient magnitude). It makes sense that if people are commenting right before the basetime they will continue to comment after.

Notice that the source of the blog post has significant terms in the interaction while the links/tracebacks do not (unless the blog post contains word 8, which occurs in only 8% of the blogs and has a very low $\beta = .014$). This implies that the features of the source of the blog post are more important than the information about the links/tracebacks as was initially believed. Contrary to another hypothesis, longer blog posts do not necessarily correspond to fewer comments, as *length* was not a significant predictor in our model (meaning we cannot conclude $\beta_{length} \neq 0$). For the same reasons, we also cannot conclude that blog posts published on Sunday or Monday have more comments. We did find that some words were significant predictors for comments in the next 24 hours. Specifically (holding all other variables constant), if word 1 appears in the blog post than we expect the number of comments in the next 24 hours to decrease by 0.118344 on average (since $\beta_{word1} = -.118344$), if word 7 appears in the blog post than we expect the number of comments in the next 24 hours to increase by 13.203045 on average (since $\beta_{word7} = 13.203045$) and if word 8 appears in the blog post than we expect the number of comments to increase by 5.123209 for each comment posted in the first 24 hours after blog post publication on average (since $\beta_{comments.first24*word8} = -5.123209$).

To improve this analysis in the future, it may be useful to collect data from blog posts with some of the words that were not found in any of the blog posts in this set of observations (words 2,3,6 and 9) to see if those have a significant effect on the number of comments.

R Code

```
# import libraries
library(MASS)
library(tree)

# source helping scripts
source("~/Desktop/36401/panelfxns.R")

# import data
data<-read.table("~/Desktop/36401/final-87.txt")
names(data)
attach(data)

n<-nrow(data)

# Univariate EDA
get.eda<-function(data){
  result<-c(min(data),max(data),median(data),mean(data),sd(data))
  return(result)
}

get.eda.cat<-function(vector,categories)
{
  index = 1
  result<-c(0)
  for (c in categories)
  {
    num<-length(which(vector==c))
    perc<-100*num/n
    result[index]=perc
    index = index + 1
  }
  return(result)
}

# print eda
get.eda(comments.next24)
get.eda(comments.prebasetime)
get.eda(comments.prev24)
get.eda(comments.first24)
get.eda(source.avg.prebasetime)
get.eda(source.avg.prev24)
get.eda(source.avg.first24)
get.eda(links.prebasetime)
get.eda(links.prev24)
get.eda(links.first24)
get.eda(length)
get.eda(post.age)
```

```
get.eda(n.parents)
get.eda(parents.ave)
get.eda(parents.min)
get.eda(parents.max)
get.eda.cat(basetime.day,c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturda
get.eda.cat(post.day,c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Saturday"))
get.eda.cat(word1,c(0,1))
get.eda.cat(word2,c(0,1))
get.eda.cat(word3,c(0,1))
get.eda.cat(word4,c(0,1))
get.eda.cat(word5,c(0,1))
get.eda.cat(word6,c(0,1))
get.eda.cat(word7,c(0,1))
get.eda.cat(word8,c(0,1))
get.eda.cat(word9,c(0,1))
get.eda.cat(word10,c(0,1))

par(mfrow=c(4,3))
hist(comments.next24,breaks=40)
hist(comments.prebasetime)
hist(comments.prev24)
hist(comments.first24)
hist(source.avg.prebasetime)
hist(source.avg.prev24)
hist(source.avg.first24)
hist(length)
hist(post.age)
hist(n.parents)
hist(parents.ave)
hist(parents.max)
title("Figure 1: Histograms for Continuous Variables",outer=TRUE)

basetime.day<-factor(basetime.day,c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday"
post.day = factor(post.day,c("Sunday","Monday","Tuesday","Wednesday","Thursday","Friday","Satu

par(mfrow=c(3,3))
title("Figure 2: Barplots for categorical variables",outer=TRUE)
barplot(table(basetime.day),main="basetime.day")
barplot(table(post.day),main="post.day")
barplot(table(word1),main="word1")
barplot(table(word4),main="word4")
barplot(table(word5),main="word5")
barplot(table(word7),main="word7")
barplot(table(word8),main="word8")
barplot(table(word10),main="word10")

par(mfrow=c(1,1))
# Multivariate EDA
vars.cont<-cbind(comments.next24,comments.prebasetime,comments.prev24,comments.first24,
```

```
            source.avg.prebasetime,source.avg.prev24,source.avg.first24,
            links.prebasetime,links.prev24,links.first24,length,post.age,
            n.parents,parents.ave,parents.max)
pairs(vars.cont,upper.panel=panel.smooth,lower.panel=panel.cor,main="Figure 3: Pairs Plots")

par(mfrow=c(2,2))
boxplot(comments.next24~basetime.day,names=c("Sunday","Monday","Tuesday","Wednesday","Thursday"
boxplot(comments.next24~post.day,names=c("Sunday","Monday","Tuesday","Wednesday","Thursday","Fr
boxplot(comments.next24~basetime.day,names=c("Sunday","Monday","Tuesday","Wednesday","Thursday"
boxplot(comments.next24~post.day,names=c("Sunday","Monday","Tuesday","Wednesday","Thursday","Fr
title("Figure 4: Categorical boxplots",outer=TRUE)

# categorical coding
Post.day<-rep(NA,length(post.day))
Post.day[post.day=="Sunday" | post.day=="Monday"] = 1
Post.day[post.day!="Sunday" & post.day!="Monday"] = 0


# weekday vs. weekend model

Basetime.day<-rep(NA,length(basetime.day))
Basetime.day[basetime.day=="Sunday" | basetime.day=="Friday" | basetime.day=="Saturday"] = 1
Basetime.day[basetime.day=="Monday" | basetime.day=="Tuesday" | basetime.day=="Wednesday" | bas

# boxplot inference model

Basetime.Day<-rep(NA,length(basetime.day))
Basetime.Day[basetime.day=="Friday" | basetime.day=="Saturday"] = 0
Basetime.Day[basetime.day=="Monday" | basetime.day=="Sunday" | basetime.day=="Wednesday" | base
Basetime.Day[basetime.day=="Tuesday"] = 2

summary(lm(comments.next24~basetime.day))
summary(lm(comments.next24~Basetime.day))
summary(lm(comments.next24~Basetime.Day))
summary(lm(comments.next24~Post.day))

# initial model
model.init<-lm(comments.next24~comments.prebasetime+comments.prev24+comments.first24+
               source.avg.prebasetime+source.avg.prev24+source.avg.first24+
               links.prebasetime+links.prev24+links.first24+length+
               post.age+Basetime.day+Post.day+word1+word4+word5+
               word7+word8+word10+n.parents+parents.ave+parents.max)

# mulitcollinearity

# partial f test
aov(lm(comments.next24~comments.prebasetime+comments.prev24+
         source.avg.prebasetime+source.avg.prev24+source.avg.first24+
         links.prebasetime+links.prev24+links.first24+length+
```

```
         post.age+Basetime.day+Post.day+word1+word4+word5+
         word7+word8+word10+n.parents+parents.ave+parents.max+comments.first24))
c.f.ssr<-15781.3
aov(lm(comments.next24~comments.prebasetime+comments.prev24+
         source.avg.prebasetime+source.avg.prev24+
         links.prebasetime+links.prev24+links.first24+length+
         post.age+Basetime.day+Post.day+word1+word4+word5+
         word7+word8+word10+n.parents+parents.ave+parents.max+comments.first24+source.avg.first
s.a.f.ssr<-1626.1
aov(lm(comments.next24~comments.prebasetime+comments.prev24+
         source.avg.prebasetime+source.avg.prev24+
         links.prebasetime+links.prev24+length+
         post.age+Basetime.day+Post.day+word1+word4+word5+
         word7+word8+word10+n.parents+parents.ave+parents.max+comments.first24+
         source.avg.first24+links.first24))
l.f.ssr<-92.2
sse<-2583093.4
f.stat.num<-(c.f.ssr+s.a.f.ssr+l.f.ssr+35357.4)/4
f.stat.den<-(sse)/(n-23)
f.stat<-f.stat.num/f.stat.den
f.stat > qf(.95,df1=3,df2=n-(22+1))
# False, therefore we can exclude these variables

model.mult1<-lm(comments.next24~comments.prebasetime+comments.prev24+
                  source.avg.prebasetime+source.avg.prev24+
                  links.prebasetime+links.prev24+length+
                  post.age+Basetime.day+Post.day+word1+word4+word5+
                  word7+word8+word10+n.parents+parents.ave+parents.max)
summary(model.mult1)


# interactions
colors<-rep(1,length(Basetime.day))
colors[Basetime.day==0]<-2
colors[Basetime.day==1]<-3
plot(comments.prev24,comments.next24,pch=16,xlab="# of Comments 24 hrs Before Basetime",
     ylab="# of Comments 24 hrs After Basetime",col=colors)
abline(lm(comments.next24[Basetime.day==0]~comments.prev24[Basetime.day==0]),col=2)
abline(lm(comments.next24[Basetime.day==1]~comments.prev24[Basetime.day==1]),col=3)
title("Figure 5: # of Comments 24 hrs After Basetime vs. # of Comments 24 hrs Before Basetime
legend("topright",c("Weekday","Weekend"),col=c(2,3),lwd=1,pch=16)
int.test1<-chisq.test(table(comments.prev24,Basetime.day))
int.test1$p.value #.5929

findInteractions<-function()
{
  pred.vars<-cbind(comments.prebasetime,comments.prev24,comments.first24,
    source.avg.prebasetime,source.avg.prev24,source.avg.first24,
    links.prebasetime,links.prev24,links.first24,length,
```

```
    post.age,Basetime.day,Post.day,word1,word4,word5,
    word7,word8,word10,n.parents,parents.ave,parents.max)
  names<-c("comments.prebasetime","comments.prev24","comments.first24","source.avg.prebasetime"
  result<-cbind()
  for (i in 1:length(names))
  {
    for (j in 1:length(names))
    {
      test<-chisq.test(table(pred.vars[i,],pred.vars[j,]))
      if (test$p.value < 4e-11)
      {
        if ((i != j) & (!(paste(names[j],names[i]) %in% result)))
        {
          result<-append(result,paste(names[i],names[j]))
        }
      }
    }
  }
  return(result)
}

findInteractions()

# lowest test p-values
c.f.l.p<-comments.first24*links.prebasetime
c.f.B.d<-comments.first24*Basetime.day
l.p.l.p24<-links.prebasetime*links.prev24
l.p.P.d<-links.prebasetime*Post.day
l.p.w8<-links.prebasetime*word8
P.d.w7<-Post.day*word7
P.d.w8<-Post.day*word8
w7.w8<-word7*word8

int.model<-lm(comments.next24~comments.prebasetime+comments.prev24+
                source.avg.prebasetime+source.avg.prev24+
                links.prebasetime+links.prev24+length+
                post.age+Basetime.day+Post.day+word1+word4+word5+
                word7+word8+word10+n.parents+parents.ave+parents.max+
                c.f.l.p+c.f.B.d+l.p.l.p24+l.p.P.d+l.p.w8+P.d.w7+
                P.d.w8+w7.w8)
summary(int.model)
aov(lm(comments.next24~comments.prebasetime+comments.prev24+
         source.avg.prebasetime+source.avg.prev24+
         links.prebasetime+links.prev24+length+
         post.age+Basetime.day+Post.day+word1+word4+word5+
         word7+word8+word10+n.parents+parents.ave+parents.max+
         c.f.l.p+l.p.l.p24+l.p.w8+
         w7.w8+P.d.w8+P.d.w7+l.p.P.d+c.f.B.d))
w7.w8.ssr<-7747.1
```

```
P.d.w8.ssr<-761.9
P.d.w7.ssr<-1140.2
l.p.P.d.ssr<-1767.7
c.f.B.d.ssr<-3719.9
sse<-2398569.1
f.num<-(w7.w8.ssr+P.d.w8.ssr+P.d.w7.ssr+l.p.P.d.ssr+c.f.B.d.ssr)/5
f.den<-sse/(n-28)
f.stat<-f.num/f.den
f.stat > qf(.95,df1=5,df2=n-(27+1))
# False, therefore we can exclude these variables from the model

model<-lm(comments.next24~comments.prebasetime+comments.prev24+
            source.avg.prebasetime+source.avg.prev24+
            links.prebasetime+links.prev24+length+
            post.age+Basetime.day+Post.day+word1+word4+word5+
            word7+word8+word10+n.parents+parents.ave+parents.max+
            c.f.l.p+l.p.l.p24+l.p.w8)

summary(model)

# remove outliers
X<-cbind(1,comments.prebasetime,comments.prev24,
           source.avg.prebasetime,source.avg.prev24,
           links.prebasetime,links.prev24,length,
           post.age,Basetime.day,Post.day,word1,word4,word5,
           word7,word8,word10,n.parents,parents.ave,parents.max,
           c.f.l.p,l.p.l.p24,l.p.w8)
H<-X%*%solve(t(X)%*%X)%*%t(X)

n<-nrow(X)
p<-ncol(X)
SSE<-sum((comments.next24-model$fitted.values)^2)
MSE<-SSE/(n-p)
res<-model$residuals; del.res<-res*sqrt((n-p-1)/(SSE*(1-diag(H))-res^2))
sort(del.res)[1:10]; sort(del.res)[(n-10):n]
alpha<-0.05; qt(1-alpha/(2*n),n-p-1) # 605, 112, 534, 50. 592, 128, 96, 312, 365
mean.h<-p/n; which(diag(H)>2*mean.h); sort(diag(H))[(n-10):n]; order(diag(H))[(n-10):n] # 442,
DFFITS<-del.res*(diag(H)/(1-diag(H)))^0.5; 2*sqrt(p/n);
sort(DFFITS[which(DFFITS > 2*sqrt(p/n))]) # 365, 96, 71, 128, 498, 442, 64, 592, 27, 436
D<-(res^2/(p*MSE))*(diag(H)/(1-diag(H))^2)
perc<-pf(D,p,n-p); tail(sort(perc)) # 365, 96, 312, 71, 82, 605
out<-c(605, 112, 534, 50, 592, 128, 96, 312, 365) # removing five observations
n<-(n-9)
comments.next24[out];order(comments.next24)[(n-10):n]


len<-length # change length to run boxcox
model.2<-lm(comments.next24[-out]+1~comments.prebasetime[-out]+comments.prev24[-out]+
              source.avg.prebasetime[-out]+source.avg.prev24[-out]+
```

```
              links.prebasetime[-out]+links.prev24[-out]+len[-out]+
              post.age[-out]+Basetime.day[-out]+Post.day[-out]+word1[-out]+word4[-out]+word5[-
              word7[-out]+word8[-out]+word10[-out]+n.parents[-out]+parents.ave[-out]+parents.ma
              c.f.l.p[-out]+l.p.l.p24[-out]+l.p.w8[-out])
summary(model.2)

par(mfrow=c(2,2))
qqnorm(model.2$res,main="Before transformation"); qqline(model.2$res)
boxcox(model.2); title("Boxcox before transformation")

model.t<-lm((comments.next24[-out]+1)^-1~comments.prebasetime[-out]+comments.prev24[-out]+
              source.avg.prebasetime[-out]+source.avg.prev24[-out]+
              links.prebasetime[-out]+links.prev24[-out]+len[-out]+
              post.age[-out]+Basetime.day[-out]+Post.day[-out]+word1[-out]+word4[-out]+word5[-
              word7[-out]+word8[-out]+word10[-out]+n.parents[-out]+parents.ave[-out]+parents.ma
              c.f.l.p[-out]+l.p.l.p24[-out]+l.p.w8[-out])
qqnorm(model.t$res,main="After transformation"); qqline(model.t$res)
boxcox(model.t); title("Boxcox after transformation")
title("Figure 6: qqnorm and boxcox",outer=TRUE)
layout(matrix(c(1,2,3,4),ncol=2))
plot(model.2$fitted,model.t$res,xlab="Fitted Values",ylab="Residuals",main="Residual plot"); ab
plot(model.t$fitted,model.t$res,xlab="Fitted Values",ylab="Residuals",main="Residual plot (tra
plot(model.2$res,pch=16,ylab="Residuals",main="Independence"); abline(h=0, lty=2)
plot(model.t$res,pch=16,ylab="Residuals",main="Independence (transform)"); abline(h=0, lty=2)
title("Figure 7: Residual plots and independence tests",outer=TRUE)
layout(matrix(c(1,2,3,4,5,6,7,8),ncol=4))
plot(word1[-out],model.2$res,ylab="Residuals",main="Residuals vs. word1"); abline(h=0, lty=2)
plot(word4[-out],model.2$res,ylab="Residuals",main="Residuals vs. word4"); abline(h=0, lty=2)
plot(word5[-out],model.2$res,ylab="Residuals",main="Residuals vs. word5"); abline(h=0, lty=2)
plot(word7[-out],model.2$res,ylab="Residuals",main="Residuals vs. word7"); abline(h=0, lty=2)
plot(word8[-out],model.2$res,ylab="Residuals",main="Residuals vs. word8"); abline(h=0, lty=2)
plot(word10[-out],model.2$res,ylab="Residuals",main="Residuals vs. word10"); abline(h=0, lty=2
plot(Post.day[-out],model.2$res,ylab="Residuals",main="Residuals vs. Post.day"); abline(h=0, l
plot(Basetime.day[-out],model.2$res,ylab="Residuals",main="Residuals vs. Basetime.day"); ablin
title("Figure 8: Residual plots for categorical variables",outer=TRUE)
layout(matrix(c(1,2,3,4,5,6,7,8,9,10,11,12,13,14),ncol=7))
plot(comments.prebasetime[-out],model.2$res,ylab="Residuals",main="Res vs. comments.prebasetim
plot(comments.prev24[-out],model.2$res,ylab="Residuals",main="Res vs. comments.prev24"); ablin
plot(source.avg.prebasetime[-out],model.2$res,ylab="Residuals",,main="Res vs. source.avg.preba
plot(source.avg.prev24[-out],model.2$res,ylab="Residuals",main="Res vs. source.avg.prev24"); ab
plot(links.prebasetime[-out],model.2$res,ylab="Residuals",main="Res vs. links.prebasetime"); ab
plot(links.prev24[-out],model.2$res,ylab="Residuals",main="Res vs. links.prev24"); abline(h=0,
plot(len[-out],model.2$res,ylab="Residuals",,main="Res vs. length"); abline(h=0, lty=2)
plot(n.parents[-out],model.2$res,ylab="Residuals",main="Res vs. n.parents"); abline(h=0, lty=2
plot(parents.ave[-out],model.2$res,ylab="Residuals",main="Res vs. parents.ave"); abline(h=0, l
plot(parents.max[-out],model.2$res,ylab="Residuals",main="Res vs. parents.max"); abline(h=0, l
plot(c.f.l.p[-out],model.2$res,ylab="Residuals",main="Res vs. comments.first24*links.prebaseti
plot(l.p.l.p24[-out],model.2$res,ylab="Residuals",main="Res vs. links.prebasetime*links.prev24
plot(l.p.w8[-out],model.2$res,ylab="Residuals",main="Res vs. links.prebasetime*word8"); abline
```

```
title("Figure 9: Residual plots for continous variables",outer=TRUE)

# Model Selection
null<-lm(comments.next24[-out]~1)
full<-model.2
model.b<-step(model.2,scope=list(lower=null,upper=model.2),method="backward")
summary(model.b)

summary(full)
aov(lm(comments.next24[-out]~comments.prebasetime[-out]+comments.prev24[-out]+
         source.avg.prebasetime[-out]+source.avg.prev24[-out]+
         post.age[-out]+word1[-out]+
         word7[-out]+c.f.l.p[-out]+l.p.w8[-out]+n.parents[-out]+
         l.p.l.p24[-out]+parents.max[-out]+parents.ave[-out]+word4[-out]+word5[-out]
        +word8[-out]+word10[-out]+Basetime.day[-out]+Post.day[-out]+len[-out]+links.prev24[-ou
n.parents.ssr<-252.15
l.p.l.p24.ssr<-242.06
parents.max.ssr<-89.01
parents.ave.ssr<-88.81
word4.ssr<-239.36
word5.ssr<-160.71
word8.ssr<-484.75
word10.ssr<-723.04
Basetime.day.ssr<-125.92
Post.day.ssr<-252.26
len.ssr<-526.53
links.prev24.ssr<-700.41
links.prebasetime.ssr<-1775.5
sse<-224004.21

f.stat.num<-(n.parents.ssr+l.p.l.p24.ssr+parents.max.ssr+parents.ave.ssr+
               word4.ssr+word5.ssr+word8.ssr+word10.ssr+Basetime.day.ssr
               +Post.day.ssr+len.ssr+links.prev24.ssr+links.prebasetime.ssr)/13
f.stat.den<-sse/(n-p)
f.stat<-f.stat.num/f.stat.den
f.stat > qf(.95,df1=13,df2=n-p)
# FALSE - we can exclude these variables from the model

model.final<-lm(comments.next24[-out]~comments.prebasetime[-out]+comments.prev24[-out]+
                  source.avg.prebasetime[-out]+source.avg.prev24[-out]+
                  post.age[-out]+word1[-out]+
                  word7[-out]+c.f.l.p[-out]+l.p.w8[-out])
summary(model.final)
```