Statistical Analysis of Bike Rents in London, UK

Julia N. Navarro

MT5762: Introductory Data Analysis

Final Project

School of Mathematics and Statistics
The University of St Andrews
St Andrews

October 2023

Abstract

The purpose of this report is to statistically investigate, for the TfL, whether and how the count of bikes rented in London, United Kingdom, depends on a variety of factors including environmental influencers, such as temperature, windspeed and humidity. The classification of days into either holidays, weekdays or workdays was also examined in the analysis.

We began by considering a subset of the data which had counts of bike rents less than 25,000 (made up around 43% of the data) and provided a 95% confidence interval associated with this approximation. We calculated the proportions of days with less than 25,000 rents in each season (e.g., 14% of the (original) data is such that it accounted for <25,000 rents in spring). A hypothesis test was then used to conclude that there was a significant difference between the proportion found between spring and winter.

For the original data set we, indeed, found changes in proportion of bike rents throughout the seasons which a Tukey test determined (largest difference was between winter and summer, as one might expect). Making use of a t-test we found that the expected number of rents between workdays and weekends does not differ significantly.

This test had a power of roughly 88% for which it would predict correctly. For a 90% accuracy we require roughly $n = 234$ observations. We lastly fitted a linear model to the data, and attempted to predict the bike count based on different predictors. R code used can be found in the Appendix.

1 Introduction

Understanding the distribution of bike rents which vary in accordance to numerous influencers, such at weather patterns and seasons, can not only help in managing the system (if repairs are required for example), but also can provide insight into unseen patterns of renting within the city. We therefore analyse this question statistically using a data set of around 100 observations which consists of the following factors: count of bike rents per day, temperature (measured in Celsius degrees), humidity (provided as a %), windspeed (given in km/h), season (spring, summer, fall and winter). Moreover, in the analysis we take into account if a day is: either a workday or not and whether it is a holiday or not.

In the next few section we will discuss the methods we used to interrogate this data set, i.e., which statistical tests were used in the process and their results. We can section our analysis into three main parts: analysis focusing on the subset of our data which contained counts of strictly less than 25,000, the analysis of the main data set, and the fitting of a linear model to predict rent count.

2 Methods and Results

We shall divide up our work into three parts. The first block of our investigation looked at the days where we had only up to 25,000 rents a day. This is easily done in R by using the filter command, and then to find the proportion we simply calculate the size of this new data set over the original. The proportion is then calculated as the ration between days with less than said counts over all days ($0.43 \equiv 43\%$). Because our data set is not representative of the population, we know that some form of uncertainty is introduced. We can therefore give a confidence interval to approximation (we choose $\alpha = 0.05$):

$$CI = (0.33, 0.53),$$

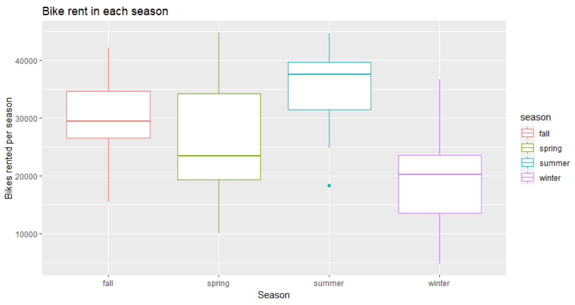
where the approximated proportion is 0.43. That is to say, equivalently, that we are 95% confident that the true proportion lies in the confidence interval.

Moreover, looking at the expected proportion (in each season) of days with <25,000 rents we can see that summer and fall have lower proportions. We calculate: 67% for spring, 2% for summer, 19% for fall and 85% for winter. This indicates that 2% of days in summer are such that less than 25,000 rents were recorded (compare this to winter where 85% of days are such like that). To find whether spring and winter proportions differ significantly we considered a hypothesis test. Our null hypothesis was that no difference exists (proportions are the same), whereas the alternative hypothesis claimed there was some difference (either larger or smaller). To perform this test we must ensure that the correct standard deviation was used (see below). Note that $p1$ and $p2$ are proportions for spring and winter, whilst $q1$ and $q2$ are given each by $q_i = 1 - p_i$, and $N2$ denotes the size of observations in the original data set (100).

$$Std_{Error} = \sqrt{\frac{(\min(p1 + p2, q1 + q2) - (p1 - p2)^2)}{N2}}.$$

Performing a t-test resulted in a value of $p = 0.0078$, which indicates that there is strong evidence to reject the null hypothesis; we accept H1 that the proportion of days differs in which rents <25,000 were recorded in.

We continue the analysis on the larger data set and examine how bike rents vary with season. To see the differences visually, we use a boxplot:



(a)

Figure 1: This plot shows the changes in bike rents in each season (fall, spring, summer and winter, and one outlier for summer). Note how small the interval is for summer compared to the other three seasons, implying overall higher bike rents.

We see that winter appears to be slightly different to the others, but we can not easily draw conclusions about the other seasons. Thus, we shall use an ANOVA test to determine whether any differences between the seasons are significant. The test yields an F-value of 23.01 for season, and using this, alongside the degrees of freedom (3 and 96), we find the associated p-value to be $2.625555 \cdot 10^{-11}$. This is essentially zero and thus indicates that at least one season has a significantly different proportion. Investigating further, we perform a Tukey test, which computes differences between means of all pairs of seasons. We indeed find that summer and winter proportions differ the most, followed by fall/winter, summer/spring proportions. This, together with the boxplot, allows us to conclude that the season in which the least bike rents occur, not surprisingly, is winter (note in the Tukey output, all other season are significanlty different when compared to winter). This is somewhat expected, that the proportions of bikes rented in winter would be lower than those, for example, in summer. People probably would consider taking the bus or perhaps the underground during these months in winter, or maybe there is another factor which causes a decrease in rents.

This next section will tackle the question of whether the type of day (workday or weekend) influences the bikes rented. The expected number of bikes rented each day is the following:

$$E_{work} = 28,622, \;\; E_{weekend} = 25,840.$$

There does appear at first glance to be a difference in number of rents, but we shall confirm this by performing a t-test. We set up our hypothesis as follows: H0: the expected number of bikes rented throughout the week is unchanging; H1: the rents do differ (either more or less during the work week). We make an assumption however: that the observations in each of the two are independent, that is that only person can only be recorded renting either on a work day or weekend. Performing a t-test, where the hypothesised difference is 0, we obtain a t-value of 1.3. Using this, and the 31 degrees of freedom, we obtain a p-value of 0.096. This therefore tells us that there is no evidence to reject the null hypothesis. Thus the expected number of rents does not depend on a working day or weekend day.

How reliable can we take this result however? To investigate this, we need to calculate the statistical power of the test (given the sample size provided and using $\alpha = 0.05$). Calculating shows that the test holds 88% power, meaning there is still a chance of detecting some event occurring even if it is not present. We can easily find the effect size required for 90% power of the test. Calculating it indicates that around 234 observations are required for said power. This overall implies that a larger sample size is required so that we do not observe a difference when there is not one in reality.

Lastly, we attempt to infer the relationship between count of bikes rented against meteorological data, season, weekday/end and holiday. We simply fit a linear model (approximation):

$$C_{rents} = \beta_0 + \beta_1 \cdot Temp. + \beta_2 \cdot Hum. + \beta_3 \cdot Wind. + \beta_4 \cdot Seas. + \beta_5 \cdot Week. + \beta_6 \cdot Hol.$$

From the output (in the table below) we can see that the temperature, humidity, weekend/day and holiday have a significant p-value associated with them. For the seasons, spring, summer and winter appear to not be the case. This just means that they do not differ much from the baseline: fall, and not that they have no relationship to rents. It would moreover appear that when the temperature increases by a degree, the bike count increases by roughly 792, whereas for example, is humidity increases by 1%, the count will drop by 295 (refer to table below for rest of slope values, where an asterisk denotes significance, where *** indicates very significant).

	Intercept	Temp.	Hum.	Wind.	Spring	Summer	Winter	Week.	Hol.
Est.	42397	792	-296	-213	-1076	-443	-1666	-3696	-8533
p-value	***	***	***	*	0.49	0.78	0.33	**	***

Table 1: Table of p-values (*, **, *** denote significance).

At this point, assuming the linear model is accurate (i.e., it satisfies assumptions: normally distributed residuals and no covariance), we can use it to predict the bike count using hypothetical scenarios. In particular, we considered the following four:

1. A work day in spring (at 18°) with 6% humidity, and 10 km/h windspeed;
2. A holiday weekend in summer (at 28°) with 35% humidity, and 5 km/h windspeed;
3. A work day in fall (at 12°) with 90% humidity, and 35 km/h windspeed;
4. A day in winter which is a weekend but not a holiday with temperature -2°C, 75% humidity, and 15 km/h windspeed.

Using the model with the appropriate values yields a table of values for the predicted count of bike rents:

Case nbr.:	fit (rents)	lower	upper
(1)	51,671	38,901	64,440
(2)	40,485	28,316	52,654
(3)	17,850	6,480	29,221
(4)	10,094	6221	13,967

Table 2: Output of the predictions

We observe how based on different inputs in our model the number of bikes rented fluctuates. Moreover, it provides upper and lower approximations to this value, with the last case, which occurs in winter, having the lowest predicted rent count. We would expect this given the conditions that the temperature is below freezing and that it is fairly windy. Given conditions like this we could get a rough idea when bike rents will peak and when they will drop, that is, playing around with the model, or using weather predictions for the next few days, could help us in predicting renting patterns.

3 Discussion

Overall, we began by investigation proportion of days between all the seasons with rents counts <25,00 and found that the lowest proportion could be found in summer and the highest in winter, as one would probably have guessed. We concluded that there is a significant difference between proportions for spring and winter.

For the original data we looked at the proportions of rents in each season and analysed them using ANOVA. This resulted in a significant p-value (indicating that at least one season differs from the others). Further investigation in R (using a Tukey test) found that the largest difference in proportion was between summer and winter. But overall the test showed that both spring and fall also differed significantly from winter.

We considered the difference in expected number of bikes rented on working days and on weekends but found no significant deviation. To ensure that this result can be trusted we calculated the power of the test and found it was at 88%. In order to increase the power from 88% to 90% we found that at least 234 observations are required.

In the last part of the analysis we fitted a linear model to our data and attempted to predict rent count. Assuming the requirements of a linear model are met, discussed below, we then used the model to predict bike rents for four scenarios.

Importantly, in the analysis, we have assumed that the residuals are normally distributed and that there was no correlation in it. In the plots below we can observe the following: the first plot appears to show evenly distributed residuals (we do not see any fanning out/in behaviour). Looking at the qq plot and the density plot, the residuals appear to be normally distributed, although the density plot appears slightly skewed to the right.

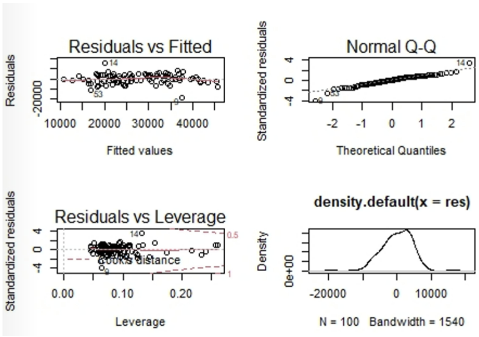


Figure 2: Plots of various types: residuals vs fitted, QQ-plot, residuals vs leverage and a density plot of the residuals

We conclude by stating that the model at hand appears to model the data reasonably, but further additions of other factors, such as time or location of bike rent could improve the usability of the model for further improved prediction use to the TfL. Moreover, additional observations would aid in improving the power of the tests performed.

To summarise, the variables, such as temperature, humidity or season, have been shown to change the outcome of number of bike rents observed. As discussed briefly, we could use the model to predict, based on current and future weather forecasts, as well as numerous other factors, such as what day it is (working day or weekend), what the expected rent count will roughly follow.

```
                           ##     Final Project MT5762    ##


## Read in data below:

Bike_sharing <-
    readr::read_csv("~/Julias_stuff/University_stuff/St_Andrews/Semester_1/MT5762
    Intro Data Analysis/Coursework/Bike_sharing.csv")


## Load in packages needed:


library(tidyverse)
library(dplyr)
library(ggplot2)
library(pwr)




## Part 2: ##

########

# a) use filter

Bike_CountLess <- Bike_sharing %>% filter(count < 25000)

##number of observations with less than 25000 observations

N1 = nrow(Bike_CountLess)

N2 = nrow(Bike_sharing)

## proportion:

P= N1/N2 #(~ 43%)


# b) use CI for proportion (95% CI => z = 1.96)

# Confidence Interval = p +/- z*sqrt( p(1-p) / n)

CIupper= P + 1.96*sqrt(P*(1-P)/N2)
CIlower= P - 1.96*sqrt(P*(1-P)/N2)

CI = c(CIlower, CIupper)
```

```
# c) the expected proportion of days with less than 25,000 rented bikes for
    each season;

Bike_Spring <- Bike_CountLess %>% filter(season == "spring")

Bike_SpringAll <- Bike_sharing %>% filter(season == "spring")

Pspring = nrow(Bike_Spring)/ nrow(Bike_SpringAll)

## ~67%

Bike_Summer <- Bike_CountLess %>% filter(season == "summer")

Bike_SummerAll <- Bike_sharing %>% filter(season == "summer")

Psummer = nrow(Bike_Summer)/ nrwo(Bike_SummerAll)

## ~2%

Bike_Fall  <- Bike_CountLess %>% filter(season == "fall")

Bike_FallAll <- Bike_sharing %>% filter(season == "fall")

Pfall = nrow(Bike_Fall)/ nrow(Bike_FallAll)

## ~19%

Bike_Winter <-Bike_CountLess %>% filter(season == "winter")

Bike_WinterAll <- Bike_sharing %>% filter(season == "winter")

Pwinter = nrow(Bike_Winter)/ nrow(Bike_WinterAll)

## ~85%

# d) difference in prop between winter/spring?

## H0 : no difference in proportions Pwinter == Pspring
## H1 : a difference in proportions Pwinter =/= Pspring

## is there a sig difference between Pspring and pwinter? need probability
    (hypthesis testing)
## test statistic =difference-hypothesised value/standard error
## case c: (participants can be in either of the four seasons:)
p1=Pspring
p2=Pwinter
q1=1-p1
q2=1-p2

min = min(p1+p2, q1+q2)

stdError = sqrt( (min - (p1-p2)^2)/N2 )

Tstat= ((p2-p1)-0)/stdError
## ~ 1.345346 = Z

# look at probability associated with said Tstat:

2*pnorm(q=Tstat, lower.tail=FALSE)

# ~ 0.007790879 => sufficient evidence to reject H0: that the proportions
    differ significantly.

#########



##  Part 3  ##

########

## Test whether the expected number of rented bikes varies across seasons.
    Interpret and explain your results.


## Do ANOVA test:


ggplot(Bike_sharing) +
  geom_boxplot(aes(x = season, y = count, col = season)) +
  ggtitle("Bike rent in each season")+
  xlab("Season")+
  ylab("Bikes rented per season")


# H0: proportions are the same for all seasons

# H1: at least one differs
# Fit ANOVA
bike.aov <- aov(count ~ season, data=Bike_sharing)
# Display ANOVA table
summary(bike.aov)

## Exact p-value
pvalue = pf(q=23.01, df1=3, df2=96, lower.tail=FALSE)

## there is difference between one or more seasons with each other.

## to find the differences between seasons, use the Tukey test:

## Tukeys HSD: gives differences between seasons:
TukeyHSD(bike.aov)

## explain the differences in seasons - clearly largest difference is between
    summer and winter.




#Furthermore, test whether there is a difference between the #### expected
    number ##### of bikes rented on working
#days and weekends. Interpret and explain your results.


## filter out data by weekend/weekday and find mean of bike rents per day:


# Data: Workdays:
Bike_Work <- Bike_sharing %>% filter (weekend == 0)

# Expected count of rents if day == working day
ExpworkCount <- mean(Bike_Work$count)


# Data: Weekends:
Bike_Weekend <- Bike_sharing %>% filter (weekend == 1)

# Expected count of rents if day == weekend
ExpweekendCount <- mean(Bike_Weekend$count)


## Our hypothesis:

## H0 : no difference in count between workday/weekend
## H1 : a difference in count between workday/weekend


## number of observations in each of the factor levels:
```

```
Owork = nrow(Bike_Work)
Oweekend = nrow(Bike_Weekend)

## sd of count in each factor level

Stdwork = sd(Bike_Work$count)
Stdweekend = sd(Bike_Weekend$count)

## calculate the standard error:

stdError2 = sqrt (Stdwork^2/Owork + Stdweekend^2/Oweekend )

###  Assumption is that observations are independent : population either rents
### in the week or on weekends (Comment to self: same with Wilcoxon test?)

Tstat2= ((ExpworkCount - ExpweekendCount)-0)/stdError2

df= min(Owork-1, Oweekend-1)

## calculate the associated probability:

pt(q = Tstat2, df = df, lower.tail = FALSE)

## -> prob ~ 0.09603715

## no evidence to reject H0: that there is no difference


########

## Power of the test:

# In addition, compute the power of the above test, assuming that the true
    difference is the one observed:

## Calculate Cohen's d:

d= (ExpweekendCount - ExpworkCount)/ sqrt( ((Stdwork)^2 + (Stdweekend)^2)/2 )

pwr.t.test(d= (ExpweekendCount - ExpworkCount)/sd(Bike_sharing$count), n= Owork
    + Oweekend ,power= NULL, sig.level=0.05, type="one.sample",
    alternative="two.sided")


## power is roughly 88%



#For the observed sample size, what effect size (i.e., difference between the
    expected values) would be
#required to obtain a power of 90%? For the given effect size, what sample size
```

```
    would be required to
#obtain a power of 90%? Explain the implications of your results for the target
    audience.



pwr.t.test(d= d, power= 0.9, sig.level=0.05, type="two.sample",
    alternative="two.sided")

## n ~ 234 in each sample (meaning in general trms that we require more data so
    that the chance of a type 2 error is 10%)


########



##  Part 4  ##

########

#Determine how the variables temperature, humidity, windspeed, season, weekend,
    and holiday affect the
#number of rented bikes. Interpret the estimated parameters of your model.

lmBike <- lm(formula = count ~ temperature + humidity + windspeed + season +
    weekend + holiday, data = Bike_sharing)

summary(lmBike)

## comment how all variables besides temperature have a negative effect on bike
    count
## largest negative influencer appears to be holiday


#Estimate the expected number
#that the TfL can expect to be rented on any given day, together with 95%
    bounds, for:


#  a) a working day in spring with temperature 18oC, 6% humidity, and 10 km/h
    windspeed;


NewData1 <- data.frame(weekend = 0, season = "spring", temperature = 18,
    humidity = 6, windspeed = 10, holiday = 0 )


predict(lmBike, newdata = NewData1, interval = "predict", level = 0.95)

#       fit      lwr       upr
#    51670.74 38900.87  64440.6


#  b) a holiday on a summer weekend with temperature 28oC, 35% humidity, and 5
   km/h windspeed;


NewData2 <- data.frame(weekend = 1, season = "summer", temperature = 28,
   humidity = 35, windspeed = 5, holiday = 1 )


predict(lmBike, newdata = NewData2, interval = "predict", level = 0.95)

#       fit     lwr       upr
#     40484.81  28315.75   52653.86


#  c) a working day in autumn with temperature 12oC, 90% humidity, and 35 km/h
   windspeed;


NewData3 <- data.frame(weekend = 0, season = "fall", temperature = 12, humidity
   = 90, windspeed = 35, holiday = 0 )


predict(lmBike, newdata = NewData3, interval = "predict", level = 0.95)

#      fit     lwr       upr
#     17850.35 6479.835 29220.86


#  d) a day on a winter weekend that is not a holiday with temperature -2oC,
   75% humidity, and 15 km/h
#windspeed.

NewData4 <- data.frame(weekend = 1, season = "winter", temperature = -2,
   humidity = 75, windspeed = 15, holiday = 0 )


predict(lmBike, newdata = NewData4, interval = "confidence", level = 0.95)

#     fit     lwr     upr
#   10093.13 6220.159 13966.1

################


## Investigate Residuals:
```

```
par(mfrow=c(3,3))

plot(lmBike)

res <- resid(lmBike)

## density plot of residuals
plot(density(res))
```