Lab Assignment 1

Guillem_Amat

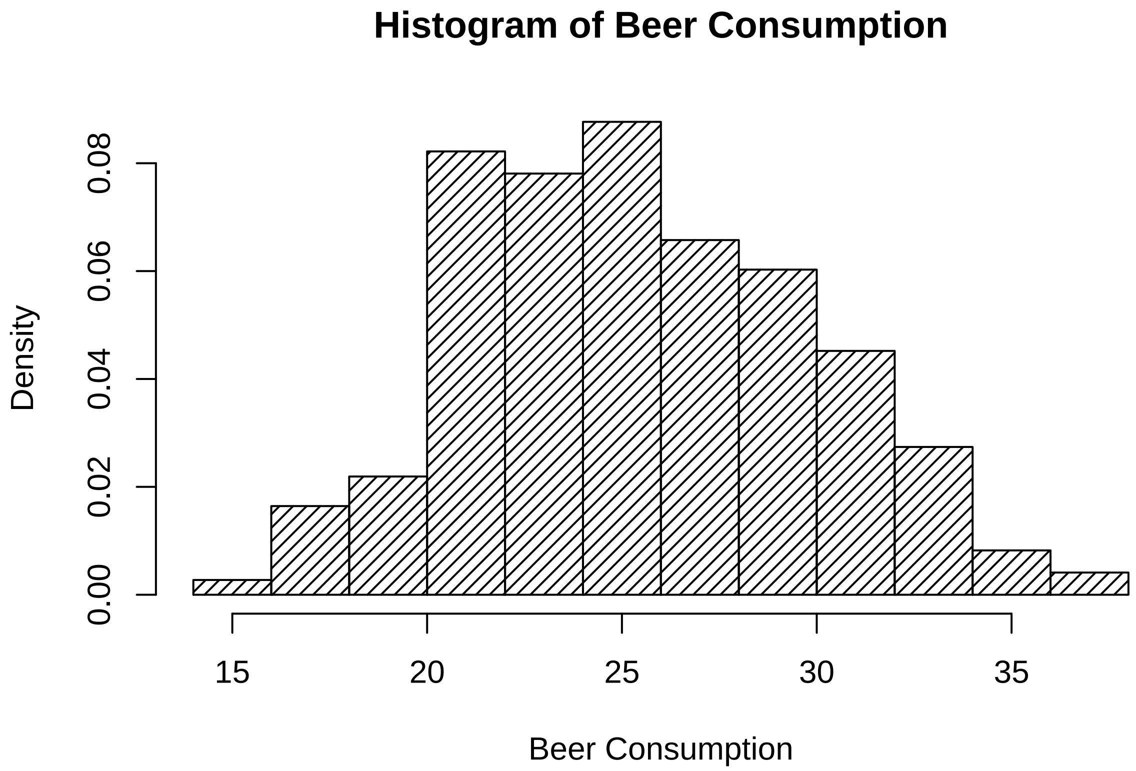
September 6, 2019

Lab Assignment 1

1. Comparing distribution

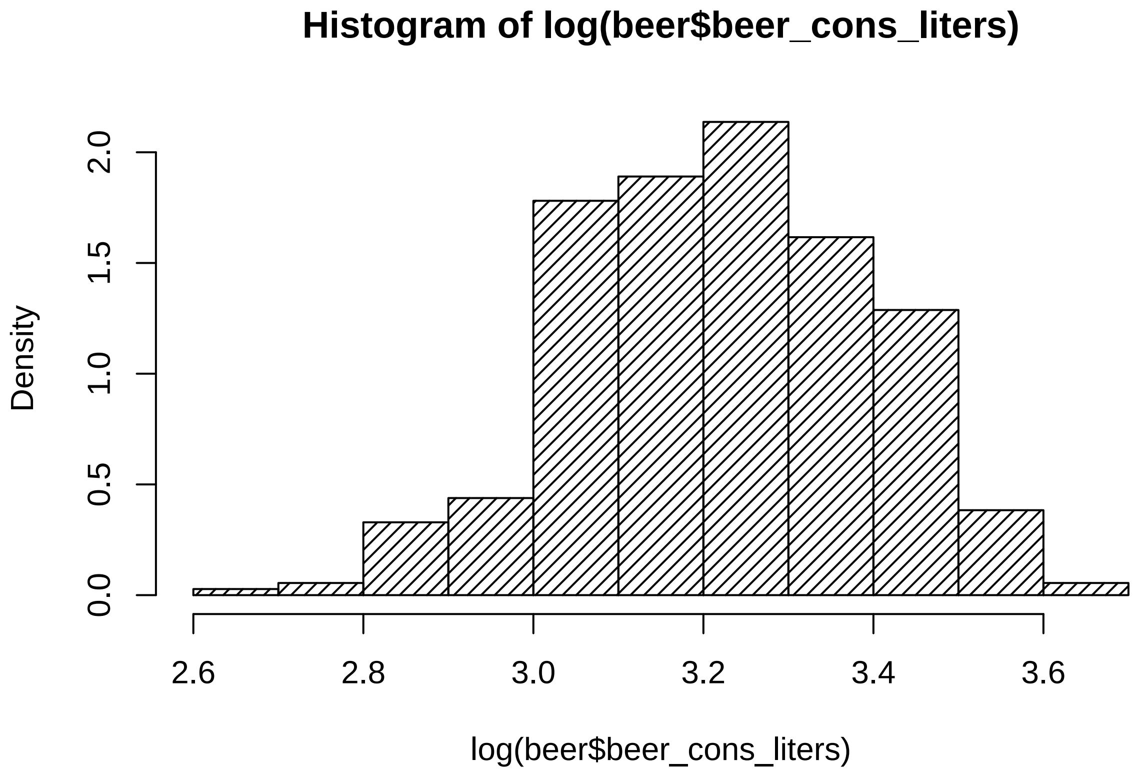
```r
#Import the data from a CSV file
beer <- read.csv(file = "C:/Users/Guillem/Desktop/Duke University/Modelling and Representation of Data/3

beer <- beer[1:365,]
```

```r
# rename the variables
beer$date <- beer$Data
beer$temp_median_c <- as.numeric(beer$Temperatura.Media..C.)
beer$temp_min_c <- as.numeric(beer$Temperatura.Minima..C.)
beer$temp_max_c <- as.numeric(beer$Temperatura.Maxima..C.)
beer$precip_mm <- as.numeric(beer$Precipitacao..mm.)
beer$weekend <- factor(beer$Final.de.Semana)
beer$beer_cons_liters <- as.numeric(beer$Consumo.de.cerveja..litros.)
beer <- beer[ , 8:ncol(beer)]

#Defining mean, standard deviation and data vector for normal distribution formula
mbeer <- mean(beer$beer_cons_liters)
sdbeer <- sqrt(var(beer$beer_cons_liters))
x <- beer$beer_cons_liters


hist(beer$beer_cons_liters, freq = FALSE, density = 20, xlab = "Beer Consumption", main = "Histogram of
```



```
#curve(dnorm(x, mbeer, sdbeer), add = TRUE, lwd = 2, col = "red")
```

```
#Trying to fit logarithmic distribution
hist(log(beer$beer_cons_liters), freq = FALSE, density = 20)
```

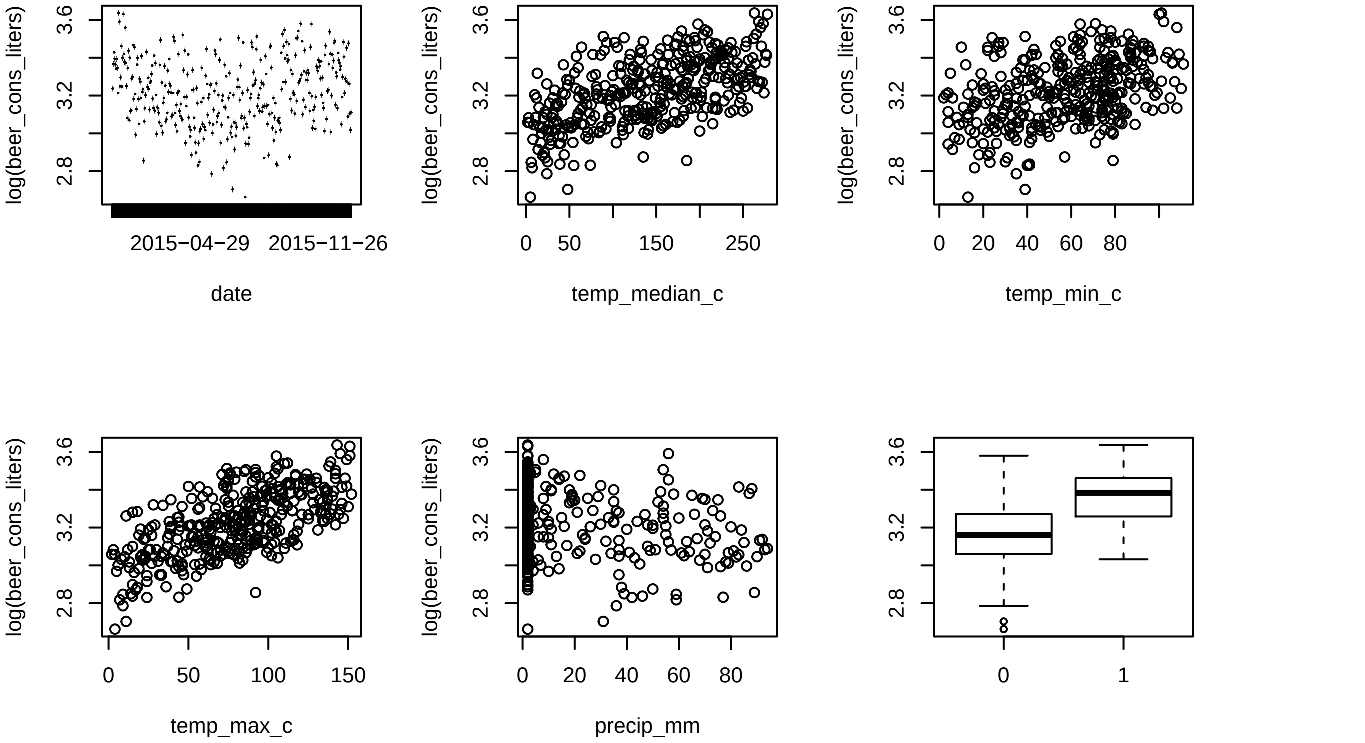


```
#The curve does not seem to work with the logarithmic histogram
#curve(dnorm(x, mbeer, sdbeer), add = TRUE, lwd = 2, col = "red")
```

It seems the non-logarithmic distribution has a more normal distribution.

2. Plots of Consumption vs Potential Predictor

```
par(mfrow = c(2, 3))

plot(log(beer_cons_liters) ~ date, data = beer)
plot(log(beer_cons_liters) ~ temp_median_c, data = beer)
plot(log(beer_cons_liters) ~ temp_min_c, data = beer)
plot(log(beer_cons_liters) ~ temp_max_c, data = beer)
plot(log(beer_cons_liters) ~ precip_mm, data = beer)
boxplot(log(beer_cons_liters) ~ weekend, data = beer)
```



Date does not seem to have a linear relationship with beer consumption, we can just observe a cloud points without a discernible pattern.

3. Temperature

```
cor(beer$temp_median_c, beer$temp_min_c)
```

```
## [1] 0.8686723
```

```
cor(beer$temp_median_c, beer$temp_max_c)
```

```
## [1] 0.9160936
```

```
cor(beer$temp_max_c, beer$temp_min_c)
```

```
## [1] 0.6744831
```

I would not say we should include all three temperature variables in our model. They are highly correlaed between each other so the model would have multicolinearity. This would make it harder for us to interpret the results.

4. Model fit

```
beer_model <- lm(beer_cons_liters ~ temp_median_c + precip_mm + weekend, data = beer)
summary(beer_model)
```

```
##
## Call:
## lm(formula = beer_cons_liters ~ temp_median_c + precip_mm + weekend,
##     data = beer)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.743 -2.096 -0.255  1.988  6.791
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   19.720943   0.321570  61.327  < 2e-16 ***
## temp_median_c  0.034486   0.001828  18.867  < 2e-16 ***
## precip_mm     -0.032696   0.005352  -6.109  2.6e-09 ***
## weekend1       5.050074   0.307009  16.449  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.641 on 361 degrees of freedom
## Multiple R-squared:  0.6426, Adjusted R-squared:  0.6396
## F-statistic: 216.3 on 3 and 361 DF,  p-value: < 2.2e-16
```

All the variables seem to have a relationship with beer consumption as indicated by the low p-values.

- Median temperature has a positive relationship with beer consumption, the hotter the temperature the more beer people drink.
- Precipitation has a negative relationship with beer consumption. The more it rains in mm the less beer people drink.
- Weekend has a positive relationship with beer consumption. On the weekends, people consume ~5 more Liters on average.

5. Least Covariate Variable

It seems that the variable that varies the least is the median temperature as it has a standard error of 0.001828.

6. Potential Limitations

There are two potential problems: Extrapolation and Correlations.

- Extrapolation is the act of predicting beyond the range of values that were used to fit the model. We can not assume that the linear trend will hold beyond our range of values.
- Correlation does not imply causality. Just because two variables vary at the same time, we can not assume that one is causing the changes in the other.

7. RMSE

```r
variables <- beer[c("weekend", "precip_mm", "temp_median_c")]
prediction <- predict(beer_model, variables)
RMSE <- mean((beer$beer_cons_liters - prediction)^2)
sqrt(RMSE)
```

```
## [1] 2.626376
```

8. K-Fold Cross Validation

```r
set.seed(25)

Kdata <- beer[sample(nrow(beer)),]
K <- 10
RMSE1 <- matrix(0,nrow=K,ncol=1)
kth_fold <- cut(seq(1,nrow(Kdata)),breaks=K,labels=FALSE)
for(k in 1:K){
  test_index <- which(kth_fold==k)
  train <- Kdata[-test_index,]
  test <- Kdata[test_index,]
  train_beer <- lm(beer_cons_liters ~ temp_median_c + precip_mm + weekend, data = train)
  Kprediction <- predict(train_beer, test)
  RMSE1[k,] <- sqrt(mean((test$beer_cons_liters- Kprediction)^2))
  }

mean(RMSE1)
```

```
## [1] 2.650413
```

9. Prediction expanded

```
beer_interaction_model <- lm(beer_cons_liters ~ temp_median_c + precip_mm + weekend + weekend:temp_media

summary(beer_interaction_model)
```

```
##
## Call:
## lm(formula = beer_cons_liters ~ temp_median_c + precip_mm + weekend +
##     weekend:temp_median_c + weekend:precip_mm, data = beer)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4182 -2.1101 -0.2415  2.0080  6.7671
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           19.658077   0.363222  54.121  < 2e-16 ***
## temp_median_c          0.034796   0.002158  16.122  < 2e-16 ***
## precip_mm             -0.031687   0.006151  -5.151 4.27e-07 ***
## weekend1               5.281708   0.673212   7.846 4.99e-14 ***
## temp_median_c:weekend1 -0.001172   0.004089  -0.287    0.775
## precip_mm:weekend1     -0.004401   0.012594  -0.349    0.727
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.647 on 359 degrees of freedom
## Multiple R-squared:  0.6428, Adjusted R-squared:  0.6378
## F-statistic: 129.2 on 5 and 359 DF,  p-value: < 2.2e-16
```

It does not seem that there is any significant interaction between weekend and the other two variables, from looking at the p-values. This can easily be interpreted: The weekend does not have any effect on temperature and precipitation.

10. RMSE on Model with Interactions

```
set.seed(25)

Kdata <- beer[sample(nrow(beer)),]
K <- 10
RSME1 <- matrix(0,nrow=K,ncol=1)
kth_fold <- cut(seq(1,nrow(Kdata)),breaks=K,labels=FALSE)
for(k in 1:K){
  test_index <- which(kth_fold==k)
  train <- Kdata[-test_index,]
  test <- Kdata[test_index,]
  train_beer <- lm(beer_cons_liters ~ temp_median_c + precip_mm + weekend + weekend:precip_mm + weekend
  Kprediction <- predict(train_beer, test)
  RMSE1[k,] <- sqrt(mean((test$beer_cons_liters- Kprediction)^2))
}

mean(RMSE1)
```

```
## [1] 2.675108
```

The RMSE for this model is slightly higher. A higher RMSE means that the data fits this model worse, as the mean difference betweeen the actual values and the predicted ones is higher.