Marketing Project

Karim Souidi

June 14, 2019

```r
options(repos=structure(c(CRAN="http://cran.utstat.utoronto.ca/")))
install.packages("magrittr")
```

```
## package 'magrittr' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\souidik\AppData\Local\Temp\Rtmp082IkN\downloaded_packages
```

```r
install.packages("tidyverse")
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\souidik\AppData\Local\Temp\Rtmp082IkN\downloaded_packages
```

```r
install.packages("ggplot2")
```

```
## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\souidik\AppData\Local\Temp\Rtmp082IkN\downloaded_packages
```

```r
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 3.5.3
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.3
```

```
## -- Attaching packages --------------------------------------------------
---------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

```
## Warning: package 'tibble' was built under R version 3.5.3
```

```
## Warning: package 'tidyr' was built under R version 3.5.3
```

```
## Warning: package 'readr' was built under R version 3.5.3
```

```
## Warning: package 'purrr' was built under R version 3.5.3

## Warning: package 'dplyr' was built under R version 3.5.3

## Warning: package 'stringr' was built under R version 3.5.3

## Warning: package 'forcats' was built under R version 3.5.3

## -- Conflicts ------------------------------------------------------------
----------------------------------- tidyverse_conflicts() --
## x tidyr::extract()   masks magrittr::extract()
## x dplyr::filter()    masks stats::filter()
## x dplyr::lag()       masks stats::lag()
## x purrr::set_names() masks magrittr::set_names()
```

```
library(ggplot2)
```

Importing the file

```
mydata<-read.csv("K:/_Staff/souidik/Bookings_data.csv", header =TRUE,
stringsAsFactors = FALSE )
```

Summary of the file

```
head(mydata)
```

```
##   ID ADR  hotel_id city_id star_rating accommadation_type_name chain_hotel
## 1  1  71   297,388   9,395         2.5                   Hotel   non-chain
## 2  2  77   298,322   9,395         3.0                   Hotel   non-chain
## 3  3 154 2,313,076   9,395         5.0                   Hotel       chain
## 4  4 127 2,240,838   9,395         3.5                   Hotel   non-chain
## 5  5 115 2,240,838   9,395         3.5                   Hotel   non-chain
## 6  6  82   331,350   9,395         3.0                   Hotel   non-chain
##   booking_date checkin_date checkout_date City time_until_checkin
## 1   02/08/2016   01/10/2016    02/10/2016    A                 60
## 2   02/08/2016   01/10/2016    02/10/2016    A                 60
## 3   02/08/2016   01/10/2016    02/10/2016    A                 60
## 4   04/08/2016   02/10/2016    03/10/2016    A                 59
## 5   04/08/2016   02/10/2016    03/10/2016    A                 59
## 6   04/08/2016   03/10/2016    05/10/2016    A                 60
##   Lengh_of_stay RevPR booking_weekend checkin_weekend dow_check_in
## 1             1    71        Week-Day        Week-End     Saturday
## 2             1    77        Week-Day        Week-End     Saturday
## 3             1   154        Week-Day        Week-End     Saturday
## 4             1   127        Week-Day        Week-End       Sunday
## 5             1   115        Week-Day        Week-End       Sunday
## 6             2   163        Week-Day        Week-Day       Monday
```

```
summary(mydata$ADR)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    63.0   114.0   148.1   192.0  3157.0
```

```
str(mydata)
```

```
## 'data.frame':    49064 obs. of  17 variables:
##  $ ID                     : chr  "1" "2" "3" "4" ...
##  $ ADR                    : int  71 77 154 127 115 82 77 168 48 82 ...
##  $ hotel_id               : chr  "297,388" "298,322" "2,313,076"
"2,240,838" ...
##  $ city_id                : chr  "9,395" "9,395" "9,395" "9,395" ...
##  $ star_rating            : num  2.5 3 5 3.5 3.5 3 3 5 2 3 ...
##  $ accommadation_type_name: chr  "Hotel" "Hotel" "Hotel" "Hotel" ...
##  $ chain_hotel            : chr  "non-chain" "non-chain" "chain" "non-
chain" ...
##  $ booking_date           : chr  "02/08/2016" "02/08/2016" "02/08/2016"
"04/08/2016" ...
##  $ checkin_date           : chr  "01/10/2016" "01/10/2016" "01/10/2016"
"02/10/2016" ...
##  $ checkout_date          : chr  "02/10/2016" "02/10/2016" "02/10/2016"
"03/10/2016" ...
##  $ City                   : chr  "A" "A" "A" "A" ...
##  $ time_until_checkin     : int  60 60 60 59 59 60 60 59 58 57 ...
##  $ Lengh_of_stay          : int  1 1 1 1 1 2 1 3 1 3 ...
##  $ RevPR                  : int  71 77 154 127 115 163 77 505 48 245 ...
##  $ booking_weekend        : chr  "Week-Day" "Week-Day" "Week-Day" "Week-
Day" ...
##  $ checkin_weekend        : chr  "Week-End" "Week-End" "Week-End" "Week-
End" ...
##  $ dow_check_in           : chr  "Saturday" "Saturday" "Saturday" "Sunday"
...
```

Formatting some variables

```r
mydata$ADR<-as.numeric(mydata$ADR)
mydata$time_until_checkin<-as.numeric(mydata$time_until_checkin)
city<-as.factor(mydata$City)
checkin_date<-as.Date(mydata$checkin_date)
booking_date<-as.Date(mydata$booking_date)
checkout_date<-as.Date(mydata$checkout_date)
```

Cheking format of Data variables

```r
class(checkin_date)
```

```
## [1] "Date"
```
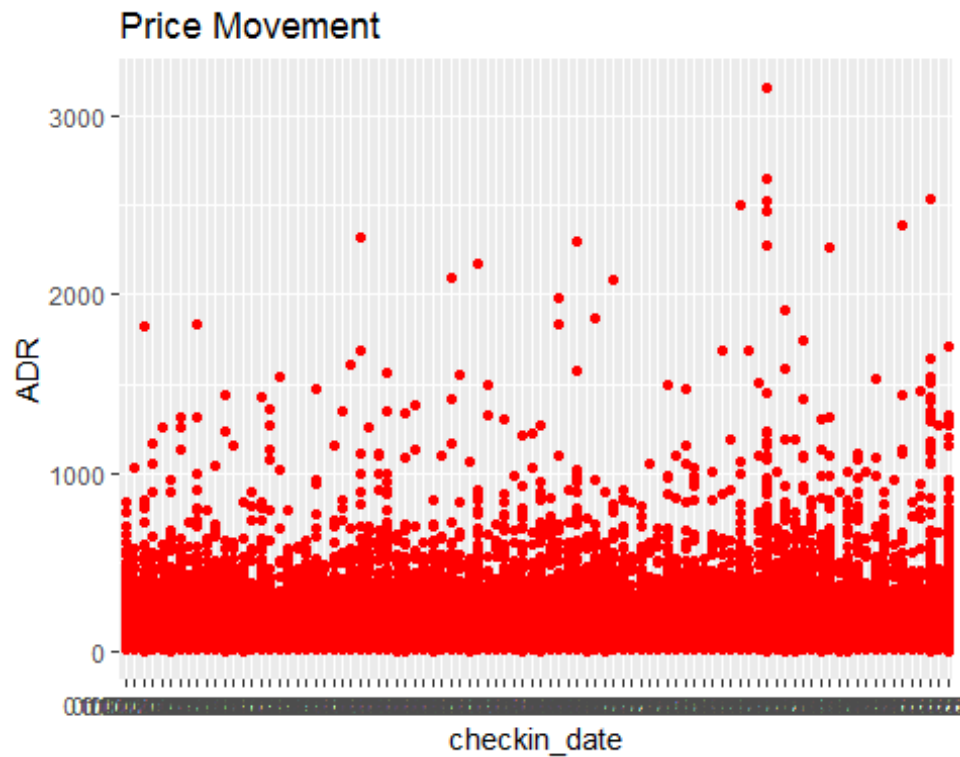
```r
class(checkin_date)
```

```
## [1] "Date"
```
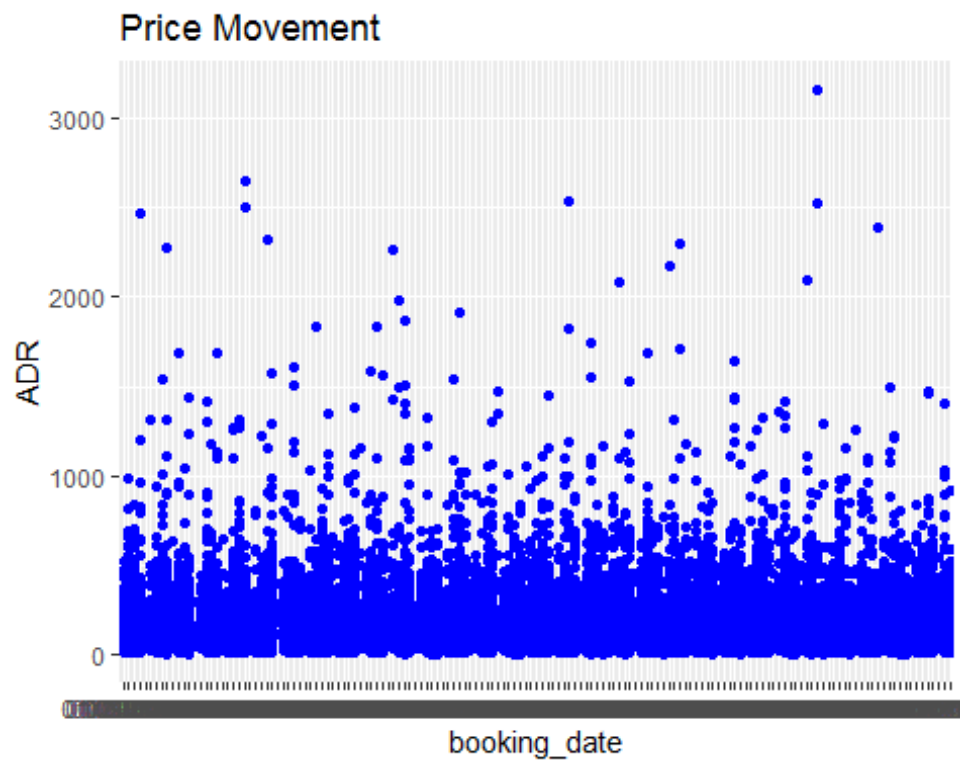
```r
class(checkin_date)
```

```
## [1] "Date"
```

```r
ggplot2::ggplot(data=mydata, aes(x = checkin_date, y = ADR)) +
        geom_point(color = "red") +
        labs(title = "Price Movement")
```

Price Movement


```r
ggplot2::ggplot(data=mydata, aes(x = booking_date, y = ADR)) +
    geom_point(color = "blue") +
    labs(title = "Price Movement")
```

Price Movement

Staring to build a demand function for hotels

```
demand<-mydata %>%
  dplyr::group_by(ADR) %>%
  dplyr::summarise(number = n())

summary(demand)

##       ADR              number
##  Min.   :   4.0   Min.   :  1.00
##  1st Qu.: 241.8   1st Qu.:  2.00
##  Median : 478.5   Median :  7.00
##  Mean   : 561.9   Mean   : 51.76
##  3rd Qu.: 740.2   3rd Qu.: 62.00
##  Max.   :3157.0   Max.   :483.00
```
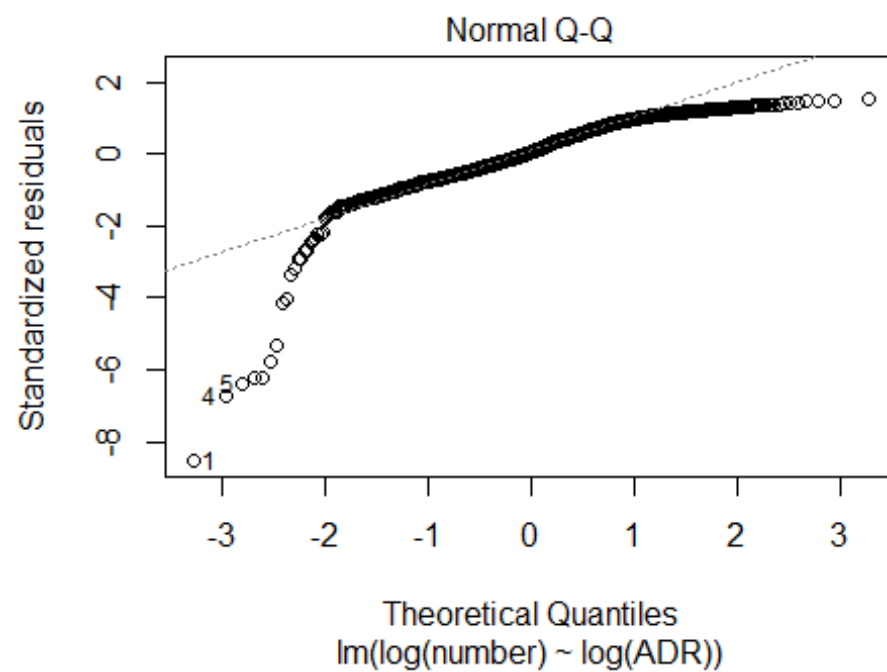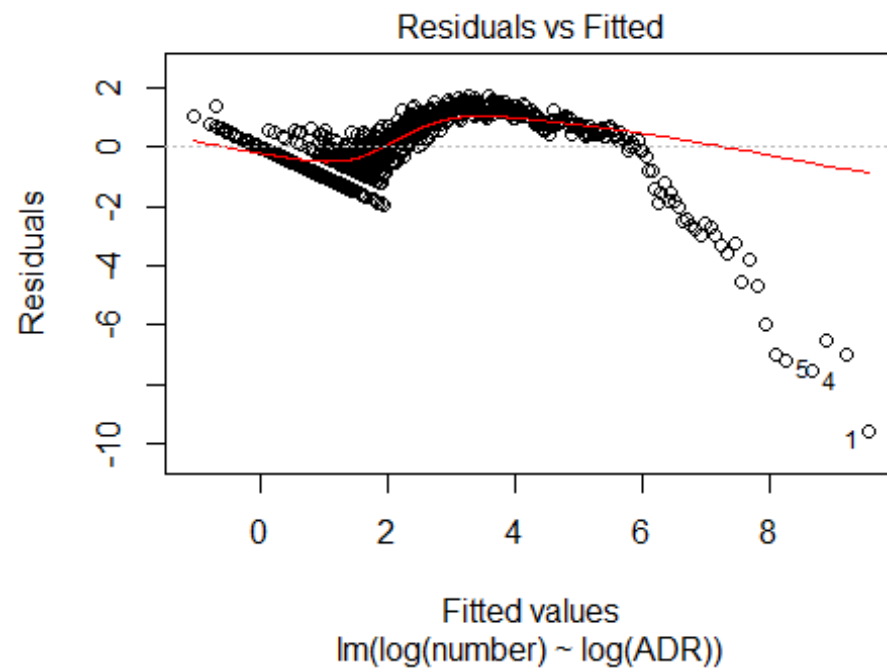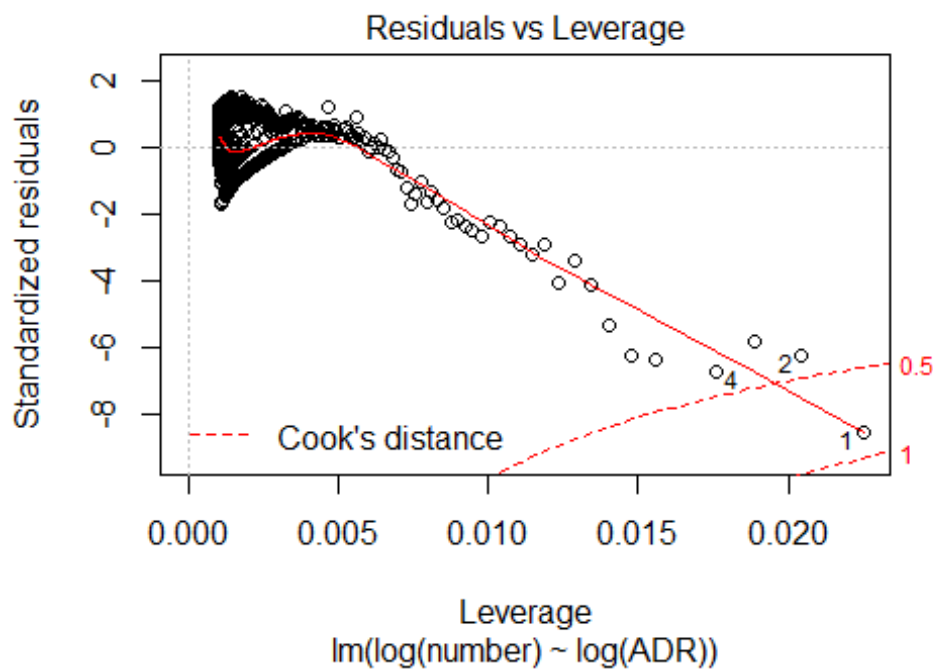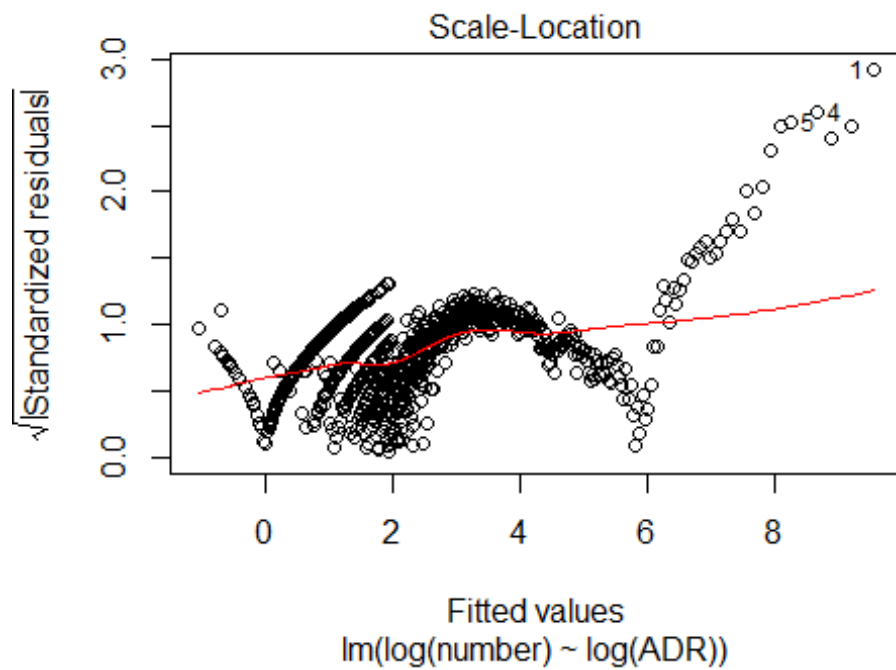
The demand function

```
demand_func<-lm(log(number)~log(ADR), data = demand)
summary(demand_func)

##
## Call:
## lm(formula = log(number) ~ log(ADR), data = demand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5515 -0.6202  0.0500  0.8261  1.7125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.75560    0.21944   53.57   <2e-16 ***
## log(ADR)    -1.58991    0.03632  -43.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.133 on 946 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6692
## F-statistic:  1917 on 1 and 946 DF,  p-value: < 2.2e-16

plot(demand_func)
```

Residuals vs Fitted

Residuals

Fitted values
lm(log(number) ~ log(ADR))

Normal Q-Q

Standardized residuals

Theoretical Quantiles
lm(log(number) ~ log(ADR))

Scale-Location

√|Standardized residuals|

Fitted values
lm(log(number) ~ log(ADR))

Residuals vs Leverage

Standardized residuals

--- Cook's distance

0.5

1

Leverage
lm(log(number) ~ log(ADR))

Dedmand function by city

```
demand_by_city <-mydata %>%
        group_by(City, ADR) %>%
```

```
        tally()
head(demand_by_city)

## # A tibble: 6 x 3
## # Groups:   City [1]
##   City    ADR     n
##   <chr> <dbl> <int>
## 1 A         4     1
## 2 A         5     9
## 3 A         6    11
## 4 A         7     2
## 5 A         9     3
## 6 A        10     1
```

Dedmand function by city Model

```
demand_by_city%>%
  filter(City == "A")%>%
  lm(log(n)~log(ADR), data =.)%>%
  summary()

##
## Call:
## lm(formula = log(n) ~ log(ADR), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7534 -0.8153  0.0961  1.1376  2.1003
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.63815    0.35807   26.92   <2e-16 ***
## log(ADR)    -1.35959    0.06761  -20.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.341 on 460 degrees of freedom
## Multiple R-squared:  0.4678, Adjusted R-squared:  0.4667
## F-statistic: 404.4 on 1 and 460 DF,  p-value: < 2.2e-16

demand_by_city%>%
  filter(City == "B")%>%
  lm(log(n)~log(ADR), data =.)%>%
  summary()

##
## Call:
## lm(formula = log(n) ~ log(ADR), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1640 -0.5023  0.0563  0.5979  1.5346
```

```
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.27585    0.21152   34.40   <2e-16 ***
## log(ADR)    -1.08529    0.03857  -28.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.8403 on 477 degrees of freedom
## Multiple R-squared:  0.624,  Adjusted R-squared:  0.6232
## F-statistic: 791.6 on 1 and 477 DF,  p-value: < 2.2e-16
```

```r
demand_by_city%>%
  filter(City == "D")%>%
  lm(log(n)~log(ADR), data =.)%>%
  summary()
```

```
## 
## Call:
## lm(formula = log(n) ~ log(ADR), data = .)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7375 -0.7443 -0.0474  0.9692  2.2710
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.29191    0.27238   26.77   <2e-16 ***
## log(ADR)    -0.94327    0.04719  -19.99   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.105 on 667 degrees of freedom
## Multiple R-squared:  0.3746, Adjusted R-squared:  0.3737
## F-statistic: 399.6 on 1 and 667 DF,  p-value: < 2.2e-16
```

```r
demand_by_city%>%
  filter(City == "E")%>%
  lm(log(n)~log(ADR), data =.)%>%
  summary()
```

```
## 
## Call:
## lm(formula = log(n) ~ log(ADR), data = .)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4376 -0.4492  0.0231  0.5719  1.6437
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    7.2090      0.2217    32.52    <2e-16 ***
## log(ADR)      -1.0501      0.0401   -26.19    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8308 on 502 degrees of freedom
## Multiple R-squared:  0.5773, Adjusted R-squared:  0.5765
## F-statistic: 685.7 on 1 and 502 DF,  p-value: < 2.2e-16
```