Advanced Data Analysis - MS4215
Predictive Modelling Assignment
ESPINOUX Jules
24267228
Code Repository: Click here to access the code

Part A: Regression modelling

First, some visualizations to see the relationship between the number of bikes hired and some other variables.

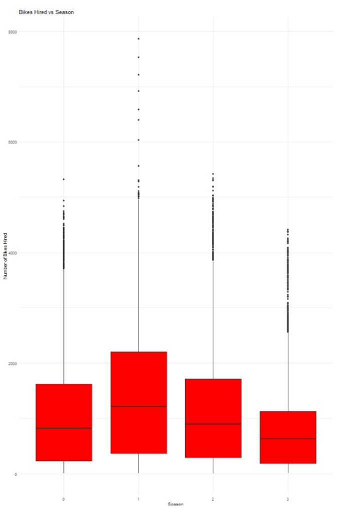


Figure 1: Bikes/Season

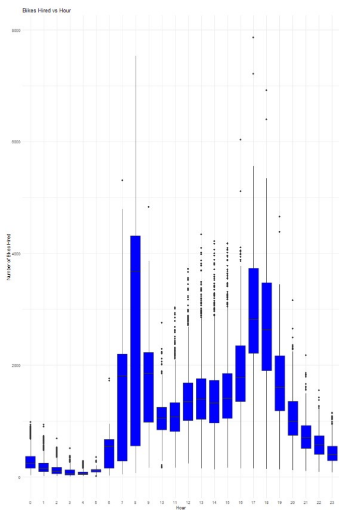


Figure 2: Bikes/Hour

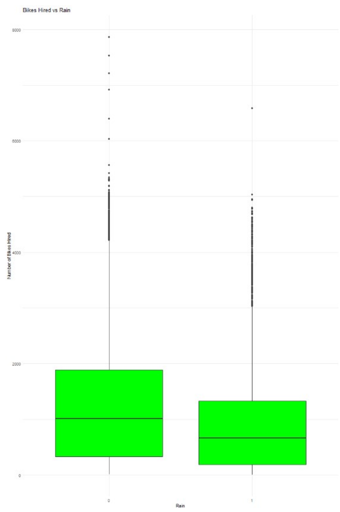
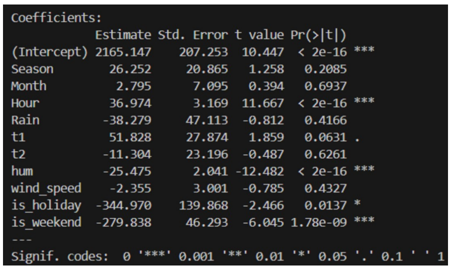


Figure 3: Bikes/Rain

We can see that there is an important dependence between the time of the day and the number of bikes hired (most bikes are hired between 7am and 8pm). We can see that Season and Rain also have an impact on the number of hired bikes (most bikes are hired during Spring and when it is not raining).

Considering all potential predictor variables, the model summary is:



```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2165.147    207.253  10.447  < 2e-16 ***
Season        26.252     20.865   1.258   0.2085
Month          2.795      7.095   0.394   0.6937
Hour          36.974      3.169  11.667  < 2e-16 ***
Rain         -38.279     47.113  -0.812   0.4166
t1            51.828     27.874   1.859   0.0631 .
t2           -11.304     23.196  -0.487   0.6261
hum          -25.475      2.041 -12.482  < 2e-16 ***
wind_speed    -2.355      3.001  -0.785   0.4327
is_holiday  -344.970    139.868  -2.466   0.0137 *
is_weekend  -279.838     46.293  -6.045 1.78e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: First model summary

Some p-values are quite low but we should check the potential multicollinearity before removing predictor variables. By printing the correlation matrix we notice that there is a high correlation between t1 and t2, this may involve multicollinearity between these two predictors.

```
                Bikes          t1          t2
Bikes     1.000000000 0.388790864 0.369021136
t1        0.388790864 1.000000000 0.988344418
```

Figure 5: Correlation matrix (part between t1 and t2)

Then we check the VIF values, with the VIF of t1 and t2 higher than 10 I make the choice to remove t2 from the potential predictor variables:

```
> vif(reg)
    Season      Month       Hour       Rain         t1         t2        hum
  1.283887   1.417891   1.138032   1.275424  59.963494  58.612327   1.984477
wind_speed is_holiday is_weekend
  1.258097   1.015856   1.012050
```

Figure 6: VIF

I then checked with anova test, the p-value is much greater than 0.05 so we can stay with this new model. Then, as the windspeed and the Month and the Rain parameters have a high p-value, I create a new model without them and I do an anova test to verify it was a good choice.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) 1745.90933  172.11208  10.144  < 2e-16 ***
Season        50.92632   19.37993   2.628 0.008660 **
Month          1.96833    6.39289   0.308 0.758195
Hour          34.71662    2.98013  11.649  < 2e-16 ***
Rain         -49.35536   44.17630  -1.117 0.264028
t1            43.45310    4.51516   9.624  < 2e-16 ***
hum          -21.82432    1.83629 -11.885  < 2e-16 ***
wind_speed     0.07167    2.69050   0.027 0.978752
is_holiday  -279.43424  139.58941  -2.002 0.045439 *
is_weekend  -159.45797   42.86019  -3.720 0.000204 ***
```

Figure 7: Second model

As the p-value is still much greater than 0.05 (0.98), I make the choice to stay with the new model. Here is the summary of the new model:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1785.385    150.627  11.853  < 2e-16 ***
Season        52.942     18.334   2.888 0.003924 **
Hour          34.487      2.969  11.616  < 2e-16 ***
t1            43.669      3.974  10.989  < 2e-16 ***
hum          -22.481      1.568 -14.337  < 2e-16 ***
is_holiday  -280.442    139.518  -2.010 0.044558 *
is_weekend  -160.855     42.825  -3.756 0.000178 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 856.9 on 1993 degrees of freedom
Multiple R-squared:  0.3313,	Adjusted R-squared:  0.3293
F-statistic: 164.6 on 6 and 1993 DF,  p-value: < 2.2e-16
```

Figure 8: Third model

When looking for the residuals, we see that a transformation is needed:

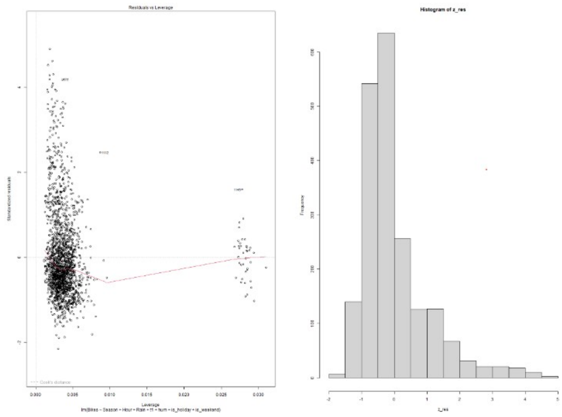


Figure 9: Residuals before transformation

With a log transformation the residuals look more normally distributed, we will stay with such a transformation for the next questions:

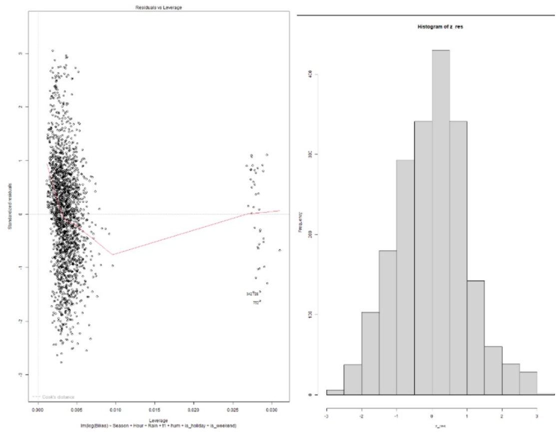


Figure 10: Residuals after transformation

After the log transformation, I see that both isholiday and isweekend parameters may not be significant:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.591723   0.165864  39.742  < 2e-16 ***
Season       0.075870   0.020189   3.758 0.000176 ***
Hour         0.081803   0.003269  25.021  < 2e-16 ***
t1           0.048349   0.004376  11.049  < 2e-16 ***
hum         -0.024879   0.001727 -14.408  < 2e-16 ***
is_holiday  -0.162357   0.153631  -1.057 0.290732
is_weekend  -0.030955   0.047157  -0.656 0.511624
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9436 on 1993 degrees of freedom
Multiple R-squared:  0.4673,    Adjusted R-squared:  0.4657
F-statistic: 291.3 on 6 and 1993 DF,  p-value: < 2.2e-16
```

Figure 11: Summary of the first log model

I remove them from the model, I do a F-TEST to check if the new model is interesting, the p-value is greater than 0.05 (0.48), I will keep the most recent model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.579078   0.165439  39.767  < 2e-16 ***
Season       0.077134   0.020158   3.826 0.000134 ***
Hour         0.081691   0.003267  25.002  < 2e-16 ***
t1           0.048639   0.004366  11.139  < 2e-16 ***
hum         -0.024927   0.001725 -14.449  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9435 on 1995 degrees of freedom
Multiple R-squared:  0.4669,    Adjusted R-squared:  0.4658
F-statistic: 436.8 on 4 and 1995 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of the model without isholiday and isweekend

I did the Cook's D diagnostic, the outliers do not have a significant impact on the model as very few points do not respect the criteria $D < 4/n$ and all the points respect the criteria $D < 1$ (with D=Cook's distance). This reinforces the choice of the model !

All the predictor variables are significant and we have the R-squared and an adjusted R-squared values quite high for a model with 2000 samples (both 0.47)

$$\log(Bikes) = 6.57 + 0.07 \cdot Season + 0.08 \cdot Hour + 0.05 \cdot t_1 - 0.02 \cdot hum \qquad (1)$$

The final model looks coherent to real life regarding the predictor variables, these variables are the most significant to me when considering hiring a bike using common sense. All of them are significant to the model, only four variables which makes the model quite 'simple', I personally do not consider many parameters when wondering if I would hire a bike !

Part B: Logistic Regression

First, let's fit a logistic regression model to the data with Attrition as the dependent variable considering all possible predictor variables in the dataset, here is the output summary:

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.472371   0.703008   2.094  0.03623 *
Age             -0.053648   0.008964  -5.985 2.17e-09 ***
BusinessTravel  -0.709922   0.139176  -5.101 3.38e-07 ***
Department       0.335204   0.139093   2.410  0.01596 *
DistanceFromHome 0.027841   0.008708   3.197  0.00139 **
Gender           0.203211   0.152373   1.334  0.18232
HourlyRate      -0.001146   0.003646  -0.314  0.75331
JobSatisfaction -0.464980   0.104614  -4.445 8.80e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1298.6  on 1469  degrees of freedom
Residual deviance: 1198.8  on 1462  degrees of freedom
AIC: 1214.8

Number of Fisher Scoring iterations: 5
```

Figure 13: Summary of the logistic regression model

Considering the output of the summary, we can see that both Gender and HourlyRate parameters are not significant to the model.

Then, I compute the odds ratios and with associated 95percent confidence intervals for each of the significant variables in the model:

```
> print(odds_ratios)
    (Intercept)             Age  BusinessTravel      Department
      4.3595603       0.9477660       0.4916826       1.3982253
DistanceFromHome          Gender      HourlyRate JobSatisfaction
      1.0282326       1.2253310       0.9988549       0.6281480
```

Figure 14: Odds ratios

```
> print(conf_intervals)
                     2.5 %     97.5 %
(Intercept)      1.1013604 17.3684564
Age              0.9309712  0.9642945
BusinessTravel   0.3736704  0.6449947
Department       1.0646973  1.8373281
DistanceFromHome 1.0107020  1.0458386
Gender           0.9109843  1.6564845
HourlyRate       0.9917368  1.0060240
JobSatisfaction  0.5110328  0.7703649
```

Figure 15: 95percent confidence intervals

I know that: if the confidence interval does not include the value 1, the effect of the variable is statistically significant. I also know that if oddsratios is greater than 1: an increase in this variable is associated with an increase of Attrition, and if oddsratios is lower than 1: an increase in this variable is associated with a reduction of Attrition.

With these elements, I can confirm that the variables Gender and HourlyRate are not significant to the model. I can also state that an increase in the department the person lives in (going from 1 to 2 or 2 to 3 or 1 to 3) and an increase in the distance from home involves an increase of the probability the person has to leave the job. Concerning the other variables (Age, Businesstravel, Job satisfaction), an increase of the value of these variables (an increase of the value representing Businesstravel means a diminution of Business travel in the job) involves a diminution of the probability the person has to leave the job.