Student Name: Ayano Yamamoto

R Version: 4.2.1

R packages required: "tidyverse", "naniar", "missMethods", "psych", "GPArotation", "rstatix", "stargazer", "car", "effectsize", "Epi", "regclass", "DescTools", "arm", "generalhoslem", "visreg"

1. Dimension Reduction

Hypotheses

- Null Hypothesis (H_0):
  Dimensions of agreeableness, extraversion, and openness cannot be reduced to a smaller set.

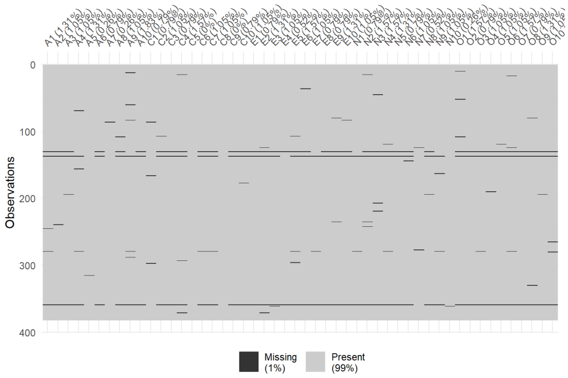
- Alternative Hypothesis (H_a):
  Dimensions of agreeableness, extraversion, and openness can be reduced to a smaller set.

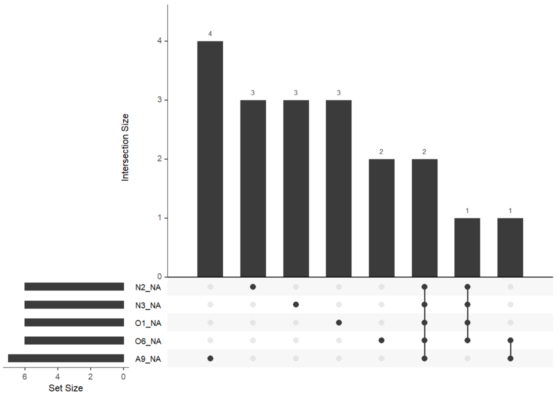
Statistical summaries

The dataset after the data preparation is a 382 x 30 data frame with no duplicates or missing values. There are no outliers with all columns having minimum values of 1 and maximum values of 5. All columns show signs of a normal distribution with skew and kurtosis within ±2.00.

To create this dataset, `studentpartII.csv` was first loaded and 30 columns related to agreeableness (`A1` - `A10`), extraversion (`E1` - `E10`), and openness (`O1` - `O10`) from the IPIP Big-Five 50-item questionnaire were selected as a subset. The sample size is 382, and there were no duplicate rows found.

Missing values were represented as 0, which have been converted to `NA`. There were 120 missing values which is 1.05% of the data frame. Visualisations of the missing values have shown that they are likely to be missing completely at random. Since there were less than 5% missing values, they have been imputed with the median values of each column.



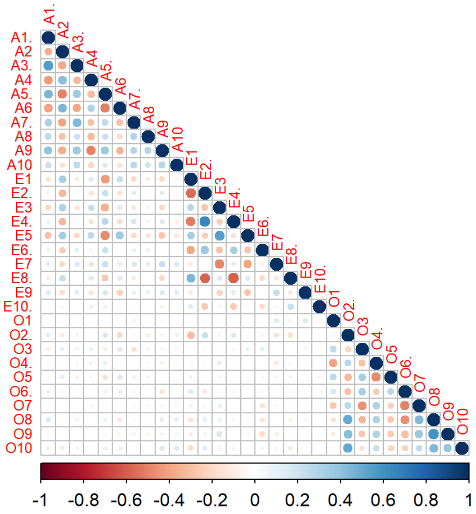


There were 12 columns from negatively worded questions, which were reverse-coded ahead of dimension reduction.

Suitability assessment

Bartlett’s test of sphericity, $\chi^2(435)$ = 3869.183, p < .001, indicated that correlations between items were sufficiently large for PCA. The KMO is greater than .6, and the determinant is greater than 0.00001 (overall MSA = 0.83, determinant = 0.000028).

Screening of the correlation matrix identified 5 variables with low correlations (`A10`, `E7`, `E9`, `E10`, `O1`) which were eliminated.

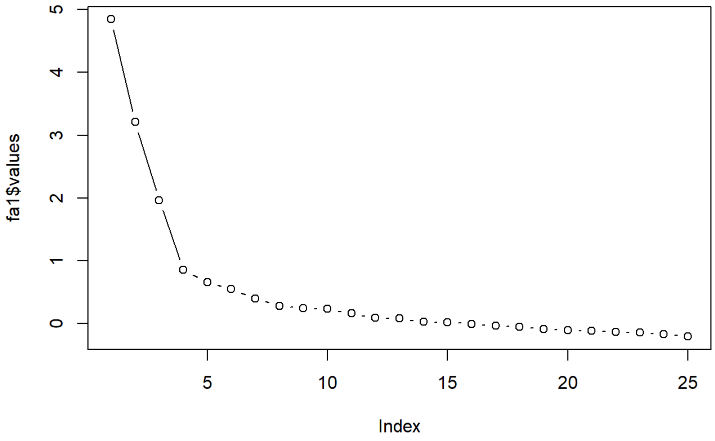


After the 5 variables have been eliminated, Bartlett’s test of sphericity, $\chi^2(435)$ = 3869.183, p < .001, still indicated that correlations between items were sufficiently large for FA. The KMO is greater than .6, and the determinant is greater than 0.00001 (overall MSA = 0.84, determinant = 0.000113).

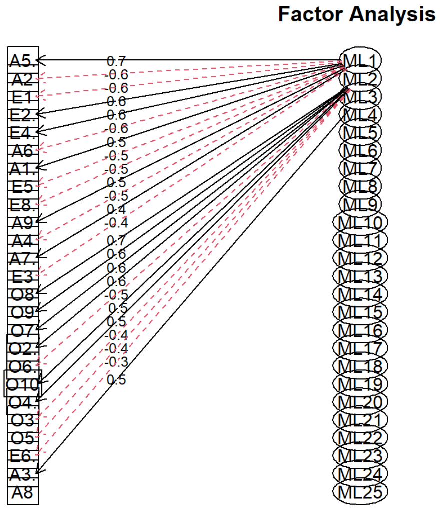
Dimension reduction

A Factor Analysis (FA) was conducted on the 30 items with Maximum Likelihood and orthogonal rotation (varimax). Factor Analysis (FA) was chosen for the dimension reduction task to achieve parsimony by explaining the maximum amount of common variance in a correlation matrix using the smallest number of explanatory constructs. Maximum Likelihood is used since the data are normally distributed as confirmed in the previous section. Varimax was selected since the potential underlying factors (agreeableness, extraversion, and openness) are independent of each other.

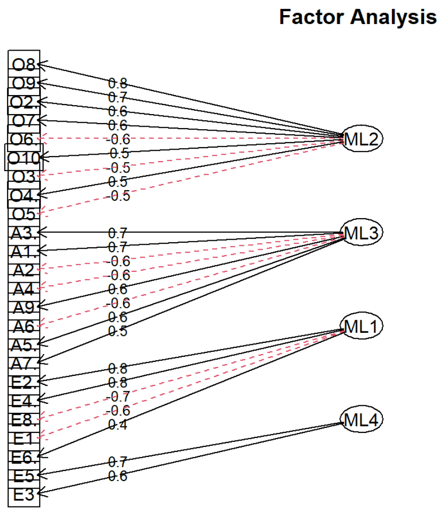
An initial analysis was run to obtain eigenvalues for each component in the data. 3 components had eigenvalues over Kaiser's criterion of 1, and in combination explained 39.13% of the variance (Kaiser, 1960). The scree plot showed inflexions that would justify retaining 4 factors. Given the large sample size, 4 components were retained in the second analysis.



Inspection of loadings and communalities showed that A8 had a communality below 0.4 and no loadings above 0.3. A8 was eliminated. There were no issues with cross-loadings.



The second analysis was run with 4 components and orthogonal rotation (varimax). Component 1 represents five questions from extroversion, component 2 openness, component 3 agreeableness, and component 4 remaining two questions from openness. 3 components had eigenvalues over Kaiser's criterion of 1 and in combination explained 37.85% of the variance. Since component 4 contains fewer than 3 items, it has been eliminated from further considerations.



Effectiveness assessment

The remaining three components, extroversion (five questions), openness, and agreeableness, all had low Cronbach's α (extroversion, Cronbach's α = -0.83; openness, Cronbach's α = 0.31; agreeableness, Cronbach's α = -0.22). After inspecting individual variables, `E1`, `E8`, `O3`, `O5`, `O6`, `A2`, `A4`, `A6` were identified as items that would improve α if removed.

The final components are extroversion (`E2`, `E4`, `E6`), openness (`O2`, `O4`, `O7`, `O8`, `O9`, `O10`), and agreeableness (`A1`, `A3`, `A5`, `A7`, `A9`). They all have acceptable reliability (extroversion, Cronbach's α = 0.71; openness, Cronbach's α = 0.79; agreeableness, Cronbach's α = 0.76) as outlined in George & Mallery (2010).

Conclusion

We will reject the null hypothesis (H_0) that the dimensions of agreeableness, extraversion, and openness cannot be reduced to a smaller set.

A Factor Analysis (FA) was conducted on the 30 items with Maximum Likelihood and orthogonal rotation (varimax). Bartlett's test of sphericity, $\chi^2(435) = 3869.183$, $p < .001$, indicated that correlations between items were sufficiently large for FA. An initial analysis was run to obtain eigenvalues for each component in the data. 3 components had eigenvalues over Kaiser's criterion of 1, and in combination explained 39.13% of the variance (Kaiser, 1960). The scree plot showed inflexions that would justify retaining 4 factors.

Given the large sample size, 4 components were retained in the second analysis, of which one contained fewer than 3 items and has been eliminated from further consideration. Component 1 represents extroversion, component 2 is openness, and component 3 is agreeableness.

All three components have acceptable reliability (extroversion, Cronbach's α = 0.71; openness, Cronbach's α = 0.79; agreeableness, Cronbach's α = 0.76).

2. Linear Regression

In sections 2 to 4, we will use the Bike Sharing dataset from part I of the CA.

Hypothesis

- Null Hypothesis (H_0):
  The temperature (actual), the weather situation, and the level of humidity cannot predict the number of bikes hired per day.

- Alternative Hypothesis (H_a):
  The temperature (actual), the weather situation, and the level of humidity can predict the number of bikes hired per day.

Statistical summaries

Outcome variable (target)

- Number of bikes hired per day

  `cnt` can be considered to follow a normal distribution (m = 4504, sd = 1937.21, n = 731, 100% of standardized scores falling within ±3.29). Therefore we can use parametric difference tests. The skewness is acceptable in order to prove normal univariate distribution, but there is a potential issue with the excess kurtosis (skewness = -.52, kurtosis = -4.48) as outlined in George & Mallery (2010).

Independent variables (predictors)

- Temperature (actual)

  `temp` can be considered to follow a normal distribution (m = .50, sd = .18, n = 731, 100% of standardized scores falling within ±3.29). The skewness is acceptable in order to prove normal univariate distribution, but there is a potential issue with the excess kurtosis (skewness = -.60, kurtosis = -6.17) as outlined in George & Mallery (2010).

  - The relationship between the total number of bikes hired per day and temperature (actual) was investigated using a Pearson correlation. A statistically significant result was found indicating a strong positive correlation (r = .63, n = 73, p < .001).

- Weather situation

  `weathersit` is a nominal categorical variable, and its groups are independent of each other. The sample size is 731, and there is no missing data. Possible values according to the dataset description are 1, 2, 3, and 4, but the dataset only includes 1, 2, and 3. The most frequently occurring value is 1, making up 63.34% of the observations.

  - Bartlett’s test was conducted to investigate the homogeneity of variance between the weather situation and the number of bikes hired per day. The p-value is 0.068, which is over 0.05, meaning that variances are homogenous in groups.
    A one-way between-groups analysis of variance (ANOVA) was conducted to explore the impact of the weather situation on the total number of bikes hired per day. There was a statistically significant difference at the $p < .05$ level in the total number of bikes hired per day for the three weather situation groups: (F(2, 728) = 40.066, $p < 0.05$. The effect size, calculated using η^2 was (0.1). Post-hoc comparisons using the Tukey HSD test indicated that the mean scores for each group differed significantly from the other two (Group 1 (M = 4876.79, SD = 1879.48), Group 2 (M = 4035.86, SD = 1809.11), Group 3 (M = 1803.29, SD = 1240.28)).

  - Bartlett’s test was conducted to investigate the homogeneity of variance between the weather situation and the temperature (actual). The p-value is 0.01, which is under 0.05, meaning that variances are heterogeneous.

    A one-way between-groups analysis of variance (ANOVA) was conducted to explore the impact of the weather situation on the temperature (actual). There was a statistically significant difference at the $p < .05$ level in the temperature (actual) for the three weather situation groups: (F(2, 57.06) = 6.49, $p < 0.05$. The effect size, calculated using η^2 was (0.19). Post-hoc comparisons using the Games Howell test indicated that the mean score for Group 1 (M = 0.51, SD = 0.19) was significantly different to that for Group 2 (M = 0.47, SD = 0.17). Group 1 was also significantly different to that for Group 3 (M = 0.43, SD = 0.13). Group 2 did not differ significantly from Group 3.

- Level of humidity

  `hum` can be considered to follow a normal distribution (m = .63, sd = .14, n = 731, 99.86% of standardized scores falling within ±3.29). Both skewness and excess kurtosis are acceptable in order to prove normal univariate distribution (skewness = -.77, kurtosis = -.36) as outlined in George & Mallery (2010).

  The minimum humidity is 0, which is an unlikely number to be observed in nature. There is 1 observation (0.14% of sample size) with a record of 0. This observation is likely to be an error and has been removed before building the predictive models.

  - The relationship between the total number of bikes hired per day and the level of humidity was investigated using a Pearson correlation. A statistically significant result was found indicating a weak negative correlation (r = -.10, n = 729, p = .01). While there is some evidence to support rejecting the null hypothesis (H_0: the level of humidity does not impact the total number of bikes hired per day) as the relationship is weak, it should be a decision made with care.

- The relationship between the temperature (actual) and the level of humidity was investigated using a Pearson correlation. A statistically significant result was found indicating a strong positive correlation (r = .13, n = 729, p < .001).

- Bartlett's test was conducted to investigate the homogeneity of variance between the weather situation and the level of humidity. The p-value is 0.00004, which is under 0.05, meaning that variances are heterogeneous.

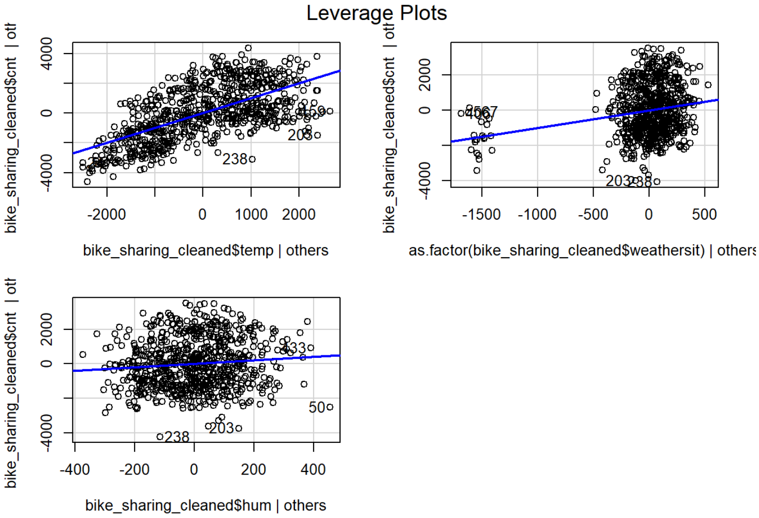
  A one-way between-groups analysis of variance (ANOVA) was conducted to explore the impact of the weather situation on the level of humidity. There was a statistically significant difference at the p < .05 level in the level of humidity for the three weather situation groups: (F(2, 52.2048285) = 6.49, p < 0.05. The effect size, calculated using η^2 was (0.87). Post-hoc comparisons using the Games Howell test indicated that the mean scores for each group differed significantly from the other two (Group 1 (M = 0.57, SD = 0.11), Group 2 (M = 0.73, SD = 0.11), Group 3 (M = 0.85, SD = 0.2)).

Regression model

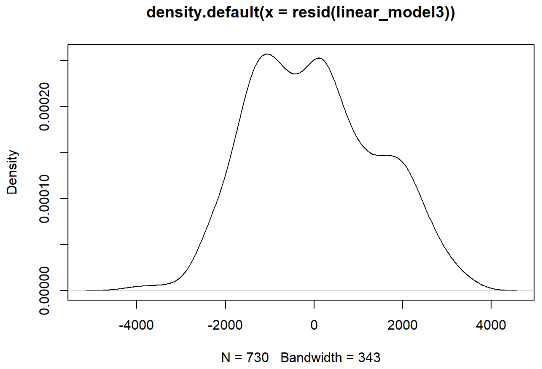
A multiple regression analysis was conducted to determine if the temperature (actual), the weather situation, and the level of humidity could predict the number of bikes hired per day. The fitted regression model was:

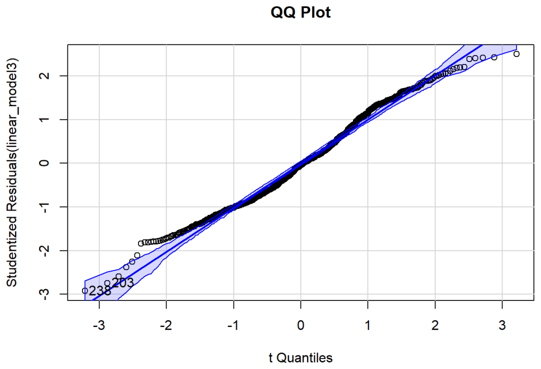
$$cnt = 2239.7 + 6551.3 * temp - 368.8 * weathersit2 - 2106.0 * weathersit3 - 1262.7 * hum$$

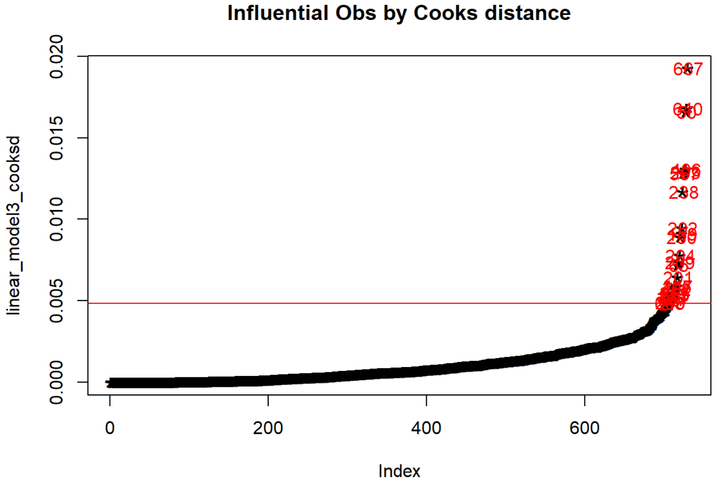
The overall regression was statistically significant (Adjusted R^2 = 0.4539, F(4, 725) = 152.5, p < .001). It was found that the temperature (actual) (`temp`) significantly predicted the number of bikes hired per day (β = 6551.3, p < .001). It was found that the weather situation (`weathersit`) of 2 significantly predicted the number of bikes hired per day (β = -368.8, p = .00872). It was found that the weather situation (`weathersit`) of 3 significantly predicted the number of bikes hired per day (β = -2106.0, p < .001). It was found that the level of humidity (`hum`) significantly predicted the number of bikes hired per day at p < 0.05 level (β = -1262.7, p = 0.01119).



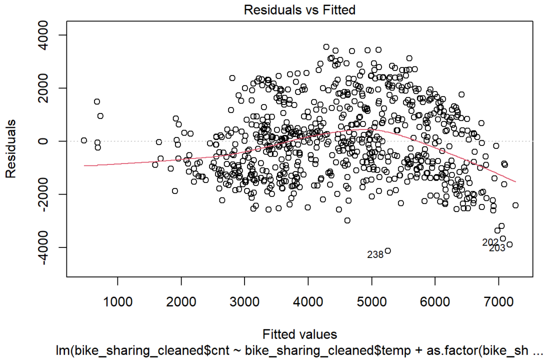
Examination of the density plot and the QQ plot of standardised residuals showed that some outliers existed. However, examination of the standardised residuals showed that none could be considered to have undue influence (Std. Residual Min = -2.91, Std. Residual Max = 2.49) and none with Cook's distance >1 as outlined in Field (2013).

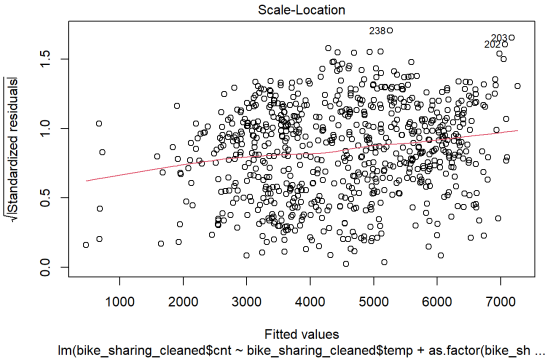






Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance > 0.4, VIF < 2.5) as outlined in Tarling (2008). The scatter plot of standardised residuals showed that the data met the assumptions of homogeneity of variance and linearity. The data also meets the assumption of non-zero variances of the predictors.





Examples from data

- The observation in row 26 has `weathersit` = 3, `temp` = .217500, `hum` = .862500, and `cnt` = 506. Using this model, the predicted value of `cnt` is $2239.7 + 6551.3 *.2175 - 2106 - 1262.7 *.8625 = 469.529$

- The observation in row 494 has `weathersit` = 2, `temp` = .575000, `hum` = .744167, and `cnt` = 4717. Using this model, the predicted value of `cnt` is $2239.7 + 6551.3 *.575 - 368.8 - 1262.7 *.744167 = 4698.237829$

- The observations in rows 8 and 14 have similar `temp` and `hum`, but different `weathersit`. `cnt` is higher for row 14.

Row	`weathersit`	`temp`	`hum`	Actual `cnt`
8	2	0.1650000	0.535833	959
14	1	0.1608700	0.537826	1421

Using this model, the predicted value of `cnt` for row 8 is $2239.7 + 6551.3 *.165 - 368.8 - 1262.7 *.535833 = 2275.268171$

The predicted value of `cnt` for row 14 is $2239.7 + 6551.3 *.16087 - 1262.7 *.537826 = 2614.494741$ which is higher than that of row 8.

Conclusion

We will reject the null hypothesis (H_0) that the temperature (actual), the weather situation, and the level of humidity cannot predict the number of bikes hired per day. The regression model explains 45.39% of the variance and is statistically significant (Adjusted $R^2 = 0.4539$, $F(4, 725) = 152.5$, $p < .001$). .

Having a higher temperature (actual) has a positive effect ($\beta = 6551.3$, $p < .001$) on the number of bikes hired per day.

Having the weather situation of 2 has a negative differential effect ($\beta = -368.8$, $p = .00872$) compared to having the weather situation of 1. Having the weather situation of 3 also has a negative differential effect ($\beta = -2106.0$, $p < .001$) compared to having the weather situation of 1, and the effect is larger than the weather situation of 2.

Having a higher level of humidity has a negative effect ($\beta = -1262.7$, $p = 0.01119$) on the number of bikes hired per day.

3. Logistic Regression

Hypothesis

- Null Hypothesis (H_0):
  The year and the temperature (actual) cannot predict the binary target of whether the number of bikes hired per day was above or below average.

- Alternative Hypothesis (H_a):
  The year and the temperature (actual) can predict the binary target of whether the number of bikes hired per day was above or below average.

Statistical summaries

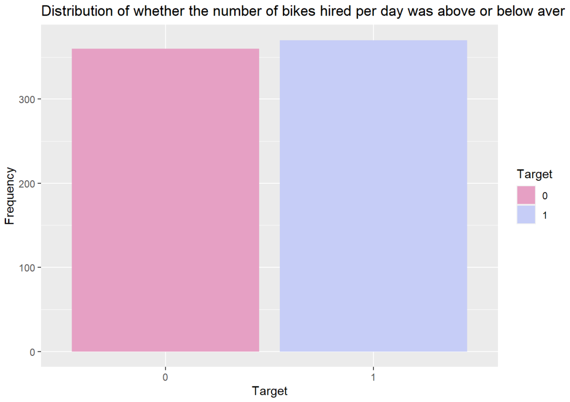
In order to include the number of bikes hired per day as a binary categorical target in the regression model, it was recorded in a new variable target (0 for below average, 1 for equal to or above average).

Outcome variable (target)

- Binary target of whether the number of bikes hired per day was above or below the average

  `target` is a binary categorical variable, and its groups are independent of each other. The sample size is 730, and there is no missing data. Possible values are 0 and 1. The most frequently occurring value is 1, making up 50.7% of the observations.

Group	Frequency	Percentage
0	360	49.3%
1	370	50.7%

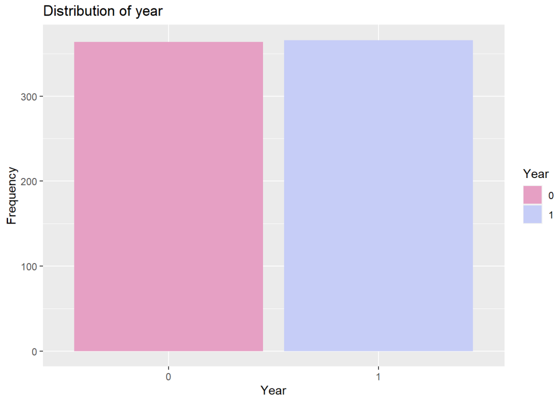


Independent variables (predictors)

- Year

  yr is a binary categorical variable, and its groups are independent of each other. The sample size is 730, and there is no missing data. Possible values are 0 and 1. The most frequently occurring value is 1, making up 50.7% of the observations.

Group	Frequency	Percentage
0	364	49.9%
1	366	50.1%



  - A Chi-square test for independence indicated a statistically significant association between year and the binary target of whether the number of bikes hired per day was above or below average, Chi2(1, n = 730) = 152.8212, p < 0.01, V = 0.4575).

- Temperature (actual)

  For the assessment of normal distribution please refer to Section 2.

  For interpretability, temp has been multiplied by 100 and stored in a new column temp_multiplied.

  - Levene's test was conducted to investigate the homogeneity of variance between the temp_multiplied and a binary target of whether the number of bikes hired per day was above or below average. The p-value is 0.00448, which is under 0.05, meaning that variances are heterogeneous in groups.

    An independent-samples t-test was conducted to compare the temp_multiplied and a binary target of whether the number of bikes hired per day was above or below average. A statistically significant difference in the temperature (M = 39.62, SD =17.16 for below average, M = 59.21, SD=13.65 for equal to or above average), (t(684.5799577)= -17.043, p < .001. Cohen's d also indicated a large effect size (-1.3).

- Levene’s test was conducted to investigate the homogeneity of variance between the `temp_multiplied` and year. The p-value is 0.07, which is over 0.05, meaning that variances are homogenous in groups.

  An independent-samples t-test was conducted to compare the `temp_multiplied` and year. No statistically significant difference in the temperature was found (M = 48.69, SD =18.98 for 2011, M = 50.41, SD=17.61 for 2012), (t(723.36)= -1.265, p = 0.21. Cohen's d also indicated a small effect size (-0.09).

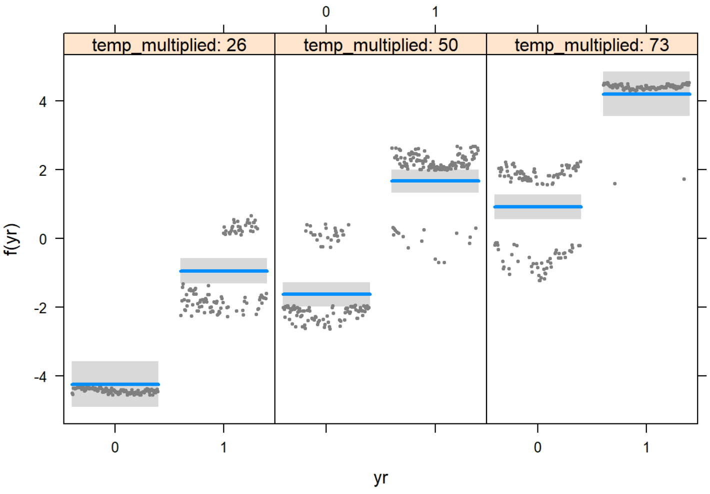
Binary logistic regression model

A multinomial logistic regression analysis was conducted with the binary target of whether the number of bikes hired per day was above or below average as the outcome variable (0 for below average, 1 for equal to or above average) with the year and the temperature (actual) as predictors.

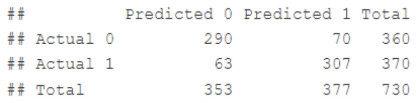
A likelihood ratio test was conducted to investigate if the model with year and temperature (actual) is an improvement over the baseline model. The improvement was found to be significant (χ^2(-2, n = 730) = 453.76, p < 0.01). Cox and Snell R^2 and Nagelkerke R^2 indicate that the model explains between 46.29% and 61.73% of the variability of whether the number of bikes hired per day was above or below average.

It was found that holding other predictor variables constant, the odds of the above-average number of bikes hired per day occurring increased by 2590.04% (95% CI [14.4276, 45.9051]) for 2012 compared to 2011. With a z-value of 11.61 and an associated p-value less than 0.01, this coefficient is statistically significant at the 5% level.

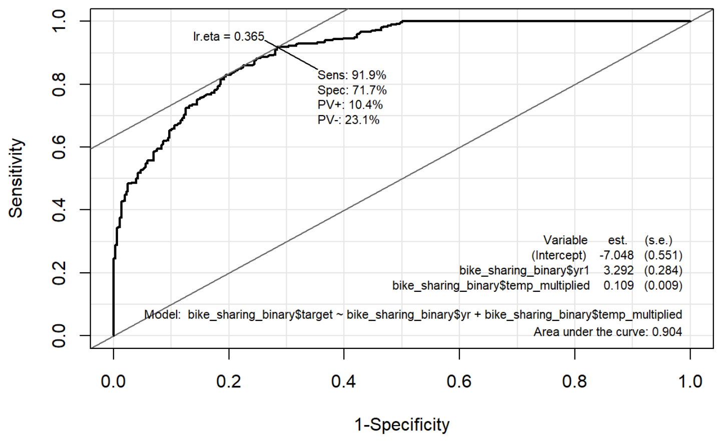
It was found that holding other predictor variables constant, the odds of the above-average number of bikes hired per day increased by 11.51% (95% CI [0.0962, 0.1342]) for each point increase in `temp_multiplied`. With a z-value of 12.54 and an associated p-value less than 0.01, this coefficient is statistically significant at the 5% level.



The confusion matrix and ROC curve show that the model performed well (sensitivity = 91.9%, specificity = 71.7%, AUC = 0.904).



```
##          Predicted 0 Predicted 1 Total
## Actual 0         290          70   360
## Actual 1          63         307   370
## Total            353         377   730
```



The data met the assumption for independent observations. Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance > 0.4, VIF < 2.5) as outlined in Tarling (2008). The Hosmer Lemeshow goodness of fit statistic did not indicate any issues with the assumption of linearity between the independent variables and the log odds of the model (χ^2(n = 8) = 11.09, p = 0.1055).

Examples from data

- The observation in row 1 has `yr` = 0, `temp_multiplied` = 34.4167, and `target` = 0. Using this model, the predicted value of `target` is 0 which is correct.

- The observation in row 500 has `yr` = 1, `temp_multiplied` = 61.1667, and `target` = 1. Using this model, the predicted value of `target` is 1 which is correct.

- The observations in rows 269 and 501 have the same `temp_multiplied` but different `yr`. `target` values are different.

Row	`yr`	`temp_multiplied`	Actual `target`
269	0	63.66670	0
501	1	63.66670	1

Using this model, the predicted values of `target` are 0 for row 269, and 1 for row 501 which are both correct.

Conclusion

We will reject the null hypothesis (H_0) that the year and the temperature (actual) cannot predict the binary target of whether the number of bikes hired per day was above or below average. The regression model explains between 46.29% and 61.73% of the variance and the improvement over the baseline model is statistically significant (χ^2(-2, n = 730) = 453.76, p < 0.01).

The year 2012 has a positive differential effect on the odds of an above-average number of bikes hired per day occurring by 2590.04% (95% CI [14.4276, 45.9051]).

Having a higher temperature (actual) has a positive effect on the odds of an above-average number of bikes hired per day occurring by 11.51% (95% CI [0.0962, 0.1342]) for each point increase in `temp_multiplied`.

4. Model Comparison

In this section we will introduce an additional model by removing a predictor (level of humidity) from the linear regression model built in section 2.

Hypothesis

- Null Hypothesis (H_0):
  The temperature (actual) and the weather situation cannot predict the number of bikes hired per day.

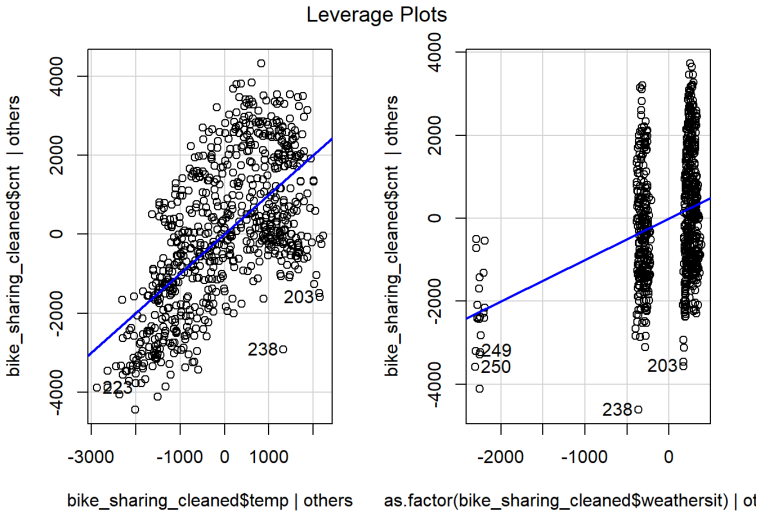
- Alternative Hypothesis (H_a):
  The temperature (actual) and the weather situation can predict the number of bikes hired per day.

Additional regression model

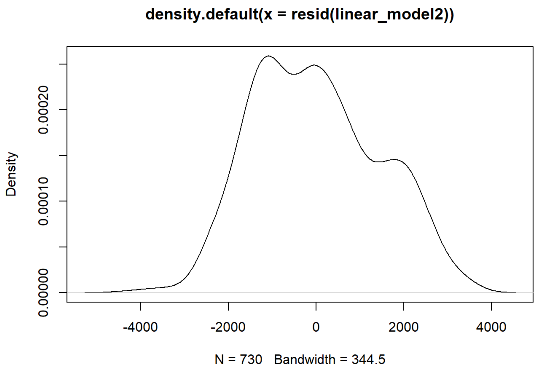
A multiple regression analysis was conducted to determine if the temperature (actual) and the weather situation could predict the number of bikes hired per day. The fitted regression model was:

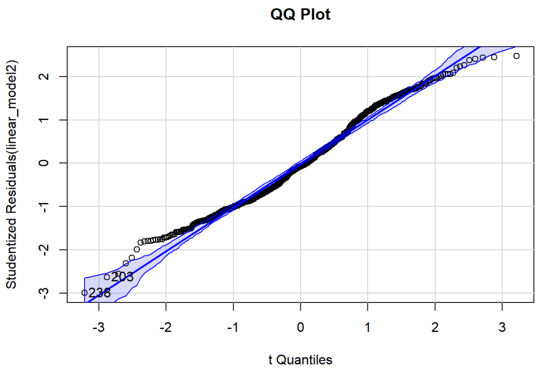
$$cnt = 1625.6 + 6355.6 * temp - 579.2 * weathersit2 - 2532.2 * weathersit3$$

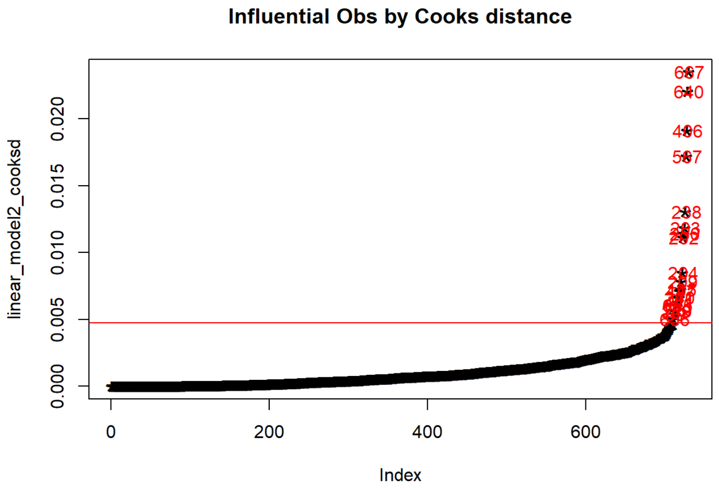
The overall regression was statistically significant (Adjusted R^2 = 0.4498, F(3, 726) = 199.7, p < .001). It was found that the temperature (actual) (`temp`) significantly predicted the number of bikes hired per day (β = 6355.6, p < .001). It was found that the weather situation (`weathersit`) of 2 significantly predicted the number of bikes hired per day (β = -579.2, p < .001). It was found that the weather situation (`weathersit`) of 3 significantly predicted the number of bikes hired per day (β = -2532.2, p < .001).



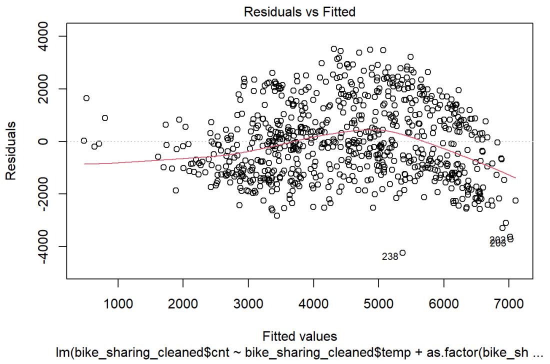
Examination of the density plot and the QQ plot of standardised residuals showed that some outliers existed. However, examination of the standardised residuals showed that none could be considered to have undue influence (Std. Residual Min = -2.97, Std. Residual Max = 2.46) and none with Cook's distance >1 as outlined in Field (2013).

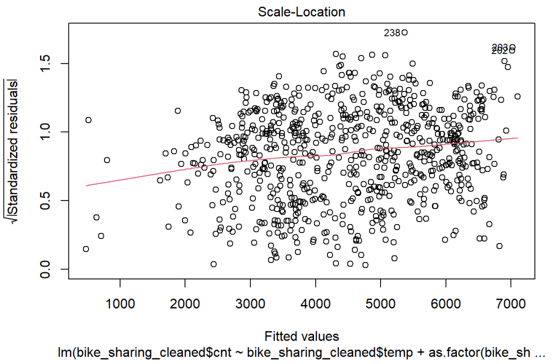






Examination for multicollinearity showed that the tolerance and variance influence factor measures were within acceptable levels (tolerance > 0.4, VIF < 2.5) as outlined in Tarling (2008). The scatter plot of standardised residuals showed that the data met the assumptions of homogeneity of variance and linearity. The data also meets the assumption of non-zero variances of the predictors.





We will reject the null hypothesis (H_0) that temperature (actual) and the weather situation cannot predict the number of bikes hired per day. The regression model explains 44.98% of the variance and is statistically significant (Adjusted R^2 = 0.4498, F(3, 726) = 199.7, p < .001).

Having a higher temperature (actual) has a positive effect (β = 6355.6, p < .001) on the number of bikes hired per day.

Having the weather situation of 2 has a negative differential effect (β = -579.2, p < .001) compared to having the weather situation of 1. Having the weather situation of 3 also has a negative differential effect (β = -2532.2, p < .001) compared to having the weather situation of 1, and the effect is larger than the weather situation of 2.

Comparison with the model from section 2

Both models are statistically significant (p < .001), and removing a predictor (the level of humidity) from the model in section 2 has decreased the adjusted R^2 only by a small amount from 0.4539 to 0.4498. The second regression model built in this section explains 0.41% less of the variance. The value of including the level of humidity as a predictor could be considered low. This was indicated in the original model, where the p-value of the level of humidity (`hum`) was the highest of all predictors (β = -1262.7, p = 0.01119).

Comparing the variables in the two models, the β value of temperature (actual) has decreased in the second model (original model, β = 6551.3; additional model, β = 6355.6). This means that having a higher temperature (actual) has a smaller positive effect on the outcome in the second model compared to the original.

The weather situation of 2 (compared to 1) has increased in β value and decreased in p-value (original model, β = -368.8, p = .00872; additional model, β = -579.2, p < .001). The β value for the weather situation of 3 (compared to 1) has decreased (original model, β = -2106.0; additional model, β = -2532.2). This indicates that the weather situation of 2 and 3 (compared to 1) have larger negative differential effects on the outcome in the second model compared to the original.

Examples from data

We will use the same examples from section 2 to compare the findings.

- The observation in row 26 has `weathersit` = 3, `temp` = .217500, `hum` = .862500, and `cnt` = 506. Using this model, the predicted value of `cnt` is
  $1625.6 + 6355.6 * .2175 - 2532.2 = 475.743$

  The predicted value from the model in section 2 was 469.529 which is further from the actual `cnt`.

- The observation in row 494 has `weathersit` = 2, `temp` = .575000, `hum` = .744167, and `cnt` = 4717. Using this model, the predicted value of `cnt` is
  $1625.6 + 6355.6 * .575 - 579.2 = 4700.87$

  The predicted value from the model in section 2 was 4698.237829 which is further from the actual `cnt`.

- The observations in rows 8 and 14 have similar `temp` and `hum`, but different `weathersit`. `cnt` is higher for row 14.

Row	`weathersit`	`temp`	`hum`	Actual `cnt`
8	2	0.1650000	0.535833	959
14	1	0.1608700	0.537826	1421

Using this model, the predicted value of `cnt` for row 8 is
$1625.6 + 6355.6 *.165 - 579.2 = 2095.074$

The predicted value for row 8 from the model in section 2 was 2275.268171 which is further from the actual `cnt`.

The predicted value of `cnt` for row 14 is
$1625.6 + 6355.6 *.16087 = 2648.025372$
which is higher than that of row 8.

The predicted value for row 14 from the model in section 2 was 2614.494741 which is closer to the actual `cnt`.

References

- Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. Sage. https://users.sussex.ac.uk/~andyf/dsusflyer.pdf
- George, D., & Mallery, P. (2010). *SPSS for Windows Step by Step: A Simple Guide and Reference*. Taylor & Francis.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and psychological measurement*, *20*(1), 141-151. https://journals.sagepub.com/doi/abs/10.1177/001316446002000116?journalCode=epma
- Tarling, R. (2008). *Statistical modelling for social researchers: Principles and practice*. Routledge. http://ndl.ethernet.edu.et/bitstream/123456789/72868/1/19pdf.pdf