IPS Instituto Politécnico de Setúbal
Escola Superior de Tecnologia de Setúbal

2º TRABALHO DE GRUPO

Métodos Estatísticos

Síntese

Neste relatório, é possível verificar toda a análise relativamente apenas à Regressão Linear Simples de um conjunto de dados previamente fornecido pelo docente responsável.

GRUPO 1:

Alexandre Coelho - 190221093

Sérgio Veríssimo - 190221128

Tim Rodrigues - 190221131

Índice

Índice de figuras

Introdução

Depois de termos constituído as observações no 1º Trabalho de Grupo, seria necessária uma análise dos dados relativamente à Regressão Linear Simples, e com isto, constituímos todo este relatório, para apresentar as análises efetuadas e todos os resultados obtidos. Relembro que o conjunto de dados permanece o mesmo, este que retrata a procura de aluguer de bicicletas na cidade de Seoul na Coreia do Sul, constituído por uma amostra de 8760 alugueis.

Descrição dos dados

As variáveis pertencentes ao conjunto de dados utilizado, já foram devidamente classificadas no relatório anterior do 1º Trabalho de Grupo. Entretanto, neste trabalho de grupo, temos a destacar apenas 3 variáveis, 2 quantitativas e 1 qualitativa, pelo que podemos relembrar de forma sucinta a descrição e classificação dessas variáveis.

Denominemos então as variáveis por variável independente (X) e variável dependente (Y), sendo que para esta análise optamos pela seguinte escolha:

$$X = Temperature(°C)(Temperatura)$$

$$Y = RentedBikeCount(Frequênciaabsolutadealugueisdebicicletasnumadata)$$

Esta decisão deve-se ao facto de conseguirmos relacionar a temperatura com o número de alugueis efetuados numa determinada data, pois quando apenas andamos de bicicleta quando as condições climatéricas são favoráveis para que tal aconteça, sendo que a temperatura é apenas um fator nessas condições climatéricas, preferimos abordar essa escolha devido a ser um dos fatores mais importantes.

Também ajudou na decisão, o facto de termos a necessidade de conseguir separar a análise ao conjunto de dados com recurso a uma variável qualitativa, que acabamos por optar pelas estações do ano (nomenclatura *Seasons* no conjunto de dados). Esta variável escolhida para a separação encaixa-se muito bem com a nossa variável independente, sendo que conseguimos avaliar também as relações que a temperatura tem com os alugueis de bicicletas nas quatro estações do ano existentes.

Regressão Linear Simples – Conjunto de dados não separado

Como nos foi requisitado a existência de uma análise relativamente à componente de Regressão Linear Simples com o conjunto de dados inteiro, e também com o conjunto de dados separado pela variável qualitativa (neste caso *Seasons* – Estações do Ano), trataremos neste tópico a análise ao conjunto de dados inteiro, não separado.

Antes de declarar que existe uma relação entre as variáveis (independente e dependente) temos de reunir as condições para verificar que essa relação existe mesmo. Para o fazermos precisamos de analisar efetivamente o conjunto de dados, com todos os passos necessários para análise de Regressão Linear Simples. Sendo assim, inicialmente, começaremos pela demonstração do resultado obtido na construção do diagrama de dispersão, que se encontra presente na Figura 1 – Diagrama de Dispersão do conjunto de dados inteiro.

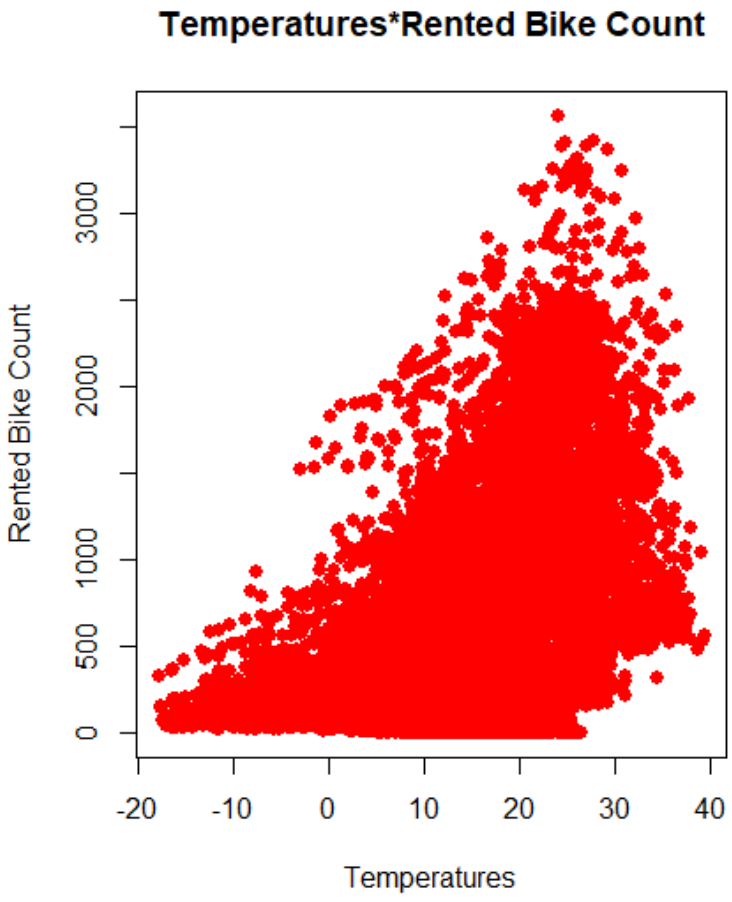

Figura 1 – Diagrama de Dispersão do conjunto de dados inteiro

Esta representação é muito útil, e conseguimos verificar uma correlação linear, quando é possível imaginar uma reta que passa pela nuvem de pontos. Aparenta existir uma correlação linear positiva, pois maiores valores de uma variável tendem a corresponder a maiores valores da outra variável. Mas não é um indicativo que devemos de seguir só pela aplicação deste método. Deveremos continuar a fazer a análise de forma a conseguirmos juntar mais provas para a potencial existência de uma relação linear entre as variáveis.

Dessa forma, é necessário posteriormente ao diagrama de dispersão, utilizar uma medida numérica que complementa a análise gráfica, e essa medida, denomina-se de **coeficiente de correlação linear de Pearson**. Esta medida indica a intensidade com que as variáveis se associam, quer positiva, quer negativamente. É representado pela seguinte equação:

$$\boldsymbol{r_{xy}} = \frac{s_{xy}}{\sqrt{s_x^2 \times s_y^2}}$$

Existem diversas maneiras de chegar ao resultado do coeficiente de correlação linear de Pearson, mas como utilizamos um comando de R próprio para obtermos o resultado final, não indicamos propriamente qual das fórmulas utilizamos, sendo que por definição, acreditamos que a equação apresentada acima, fora aplicada no código da biblioteca do comando.

Com a execução desse comando, chegamos ao resultado:

$$\boldsymbol{r_{xy}} = 0.5385582$$

Este resultado obtido indica que estamos perante um coeficiente de correlação linear adequado, pois encontra-se dentro dos seus parâmetros:

$$\boldsymbol{-1 \leq r_{xy} \leq 1}$$

Também indica que estamos perante uma correlação linear positiva, que poderia ser confirmada a olho, na análise gráfica ao diagrama de dispersão anteriormente referido.

Relativamente à intensidade da correlação, tendo como base a Figura 2 - Tabela indicativa da intensidade da correlação, podemos afirmar que estamos perante uma correlação linear positiva moderada, pois:

$$r_{xy} = 0.5385582 \in [0.5,0.8[$$

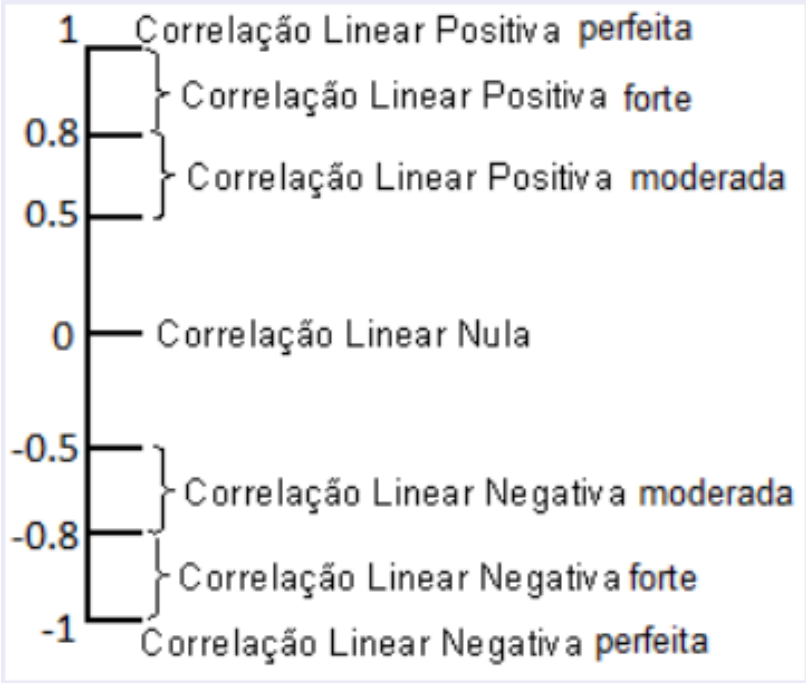

Figura 2 – Tabela indicativa da intensidade da correlação

Mas como tal, ainda não podemos afirmar que a análise se encontra devidamente efetuada, e que podemos afirmar que existe uma relação entre as duas variáveis adequada, pois falta-nos efetuar uma análise residual ao conjunto de dados, mas para tal é necessário determinar a reta de regressão, que se rege pela seguinte equação:

$$\hat{y} = a + bx$$

Esta reta, quando existe uma correlação entre as duas variáveis, permite-nos efetuar previsões para ambas as variáveis.

Relativamente à equação, afirmamos que:

- A "incógnita" **a** – Representa a ordenada na origem, ou seja, o valor de y que se espera observar quando x = 0. Ou podemos indicar por outros termos, que é o local onde a reta corta o eixo dos yy;
- A "incógnita" **b** – Representa o declive, ou inclinação da reta. O valor indica em que medida y muda em função de x, refletindo a correlação existente as variáveis. Ou seja, quando b é positivo, existe uma correlação linear positiva, quando b é negativo, existe uma correlação linear negativa.

Para calcularmos estas duas "incógnitas", utilizamos fórmulas próprias para obter um valor, estas fórmulas são:

$$\boldsymbol{a} = \bar{y} - b\bar{x}$$

$$\boldsymbol{b} = \frac{s_{xy}}{s_x^2}$$

Para efetivamente determinamos os valores da reta, recorremos a um comando R que tem o propósito de efetuar exatamente esse cálculo. Esse comando retornou-nos a seguinte reta:

$$\hat{\boldsymbol{y}} = \mathbf{329.95} + \mathbf{29.08}\boldsymbol{x}$$

Como tal, não podemos afirmar ainda nada relevante relativamente a este método aplicado, mas tal pode ser utilizado para o cálculo dos resíduos e para demonstrar que a relação linear entre as duas variáveis é adequada. Entretanto, também nos permite agora, efetuar uma demonstração do diagrama de dispersão com a reta de regressão obtida. Este diagrama de dispersão pode ser consultado através da Figura 3 – Diagrama de Dispersão do conjunto de dados inteiros com reta de regressão.

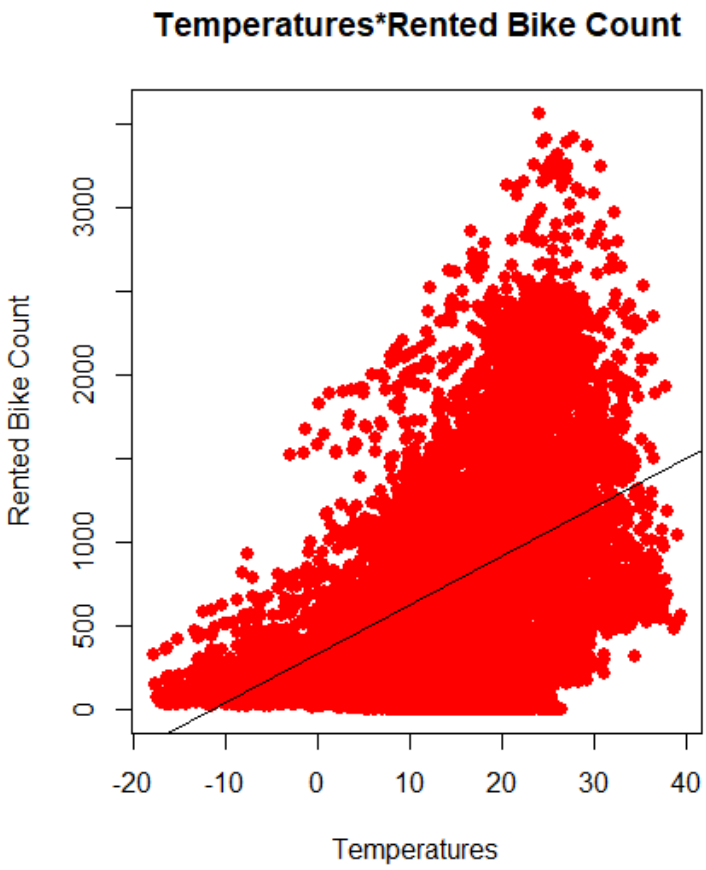

Figura 3 - Diagrama de Dispersão do conjunto de dados inteiros com reta de regressão

Como mencionado anteriormente, também podemos fazer algumas previsões com recurso à reta de regressão que obtivemos. Referindo também, que esta previsões podem não se demonstrar adequadas, dependendo assim, da análise completa a este modelo.

Como por exemplo, quantas bicicletas serão alugadas se tiver uma temperatura de 30℃ numa determinada data:

$$\hat{y}(30) = 329.95 + 29.08 \times 30 = 1202.35$$

Ou quantas bicicletas serão alugadas se tiver uma temperatura de -10℃ numa determinada data:

$$\hat{y}(-10) = 329.95 + 29.08 \times -10 = 39.15$$

Entretanto, existem algumas previsões que incorrem numa previsão absurda, como por exemplo, se existir um dia com temperatura de -15ºC, obtemos o seguinte resultado:

$$\hat{\boldsymbol{y}}(\mathbf{-15}) = \mathbf{329.95} + \mathbf{29.08} \times \mathbf{-15} = \mathbf{-106.25}$$

O que indica que se existir uma temperatura de -15ºC, que é comum ocorrer no inverno em Seoul, temos um aluguel de bicicletas de -106.25, o que é impossível, pois não existem alugueis negativos.

Relativamente à previsão de temperatura através do número de bicicletas alugadas numa determinada data, teremos que recorrer à seguinte equação:

$$\hat{\boldsymbol{x}} = a^* + b^* y$$

Sendo o nosso a^* e b^*, respetivamente calculados com as seguintes fórmulas:

$$a^* = \bar{x} - b^* \bar{y}$$

$$b^* = r_{xy} \times \frac{s_x}{s_y}$$

Com recurso ao comando R anteriormente utilizado, mas trocando as variáveis de parâmetros, obtemos o seguinte resultado:

$$\hat{\boldsymbol{x}} = 5.855464 + 0.009974y$$

Com isto podemos fazer algumas previsões da temperatura através do número de bicicletas, como por exemplo, que temperatura estaria, se fossem alugadas 500 bicicletas:

$$\hat{\boldsymbol{x}}(500) = 5.855464 + 0.009974 \times 500 = 10.842464$$

Ou caso existisse apenas 1 aluguel de uma bicicleta:

$$\hat{\boldsymbol{x}}(1) = 5.855464 + 0.009974 \times 1 = 5.865438$$

Entretanto, não podemos afirmar que estas previsões se encontrem muito corretas. Podemos verificar isso, caso tentemos prever com os valores extremos de alugueis de bicicletas, como por exemplo, o valor máximo determinado e apresentado no relatório anterior de número de bicicletas alugadas num determinado dia, que fora 3556 bicicletas:

$$\hat{\boldsymbol{x}}(1) = 5.855464 + 0.009974 \times 3556 = 41.323008$$

Embora desse uma temperatura válida, tal não se pode considerar, pois o valor 41.323008, cede o valor máximo que fora determinado e apresentado no relatório anterior relativamente à temperatura, que era 39.4℃.

Ou seja, é absurdo pois:

$$41.323008 > 39.4$$

Não pertencendo, portanto, ao intervalo de valores observados.

Relativamente aos resíduos, obtivemos o gráfico residual presente na Figura 4 – Gráfico de resíduos do conjunto de dados inteiro, que indica que não existe um padrão bem definido, portanto o modelo ajustado parece ser adequado. Embora o modelo tenha sido considerado adequado (diagrama de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados, criando assim uma previsão com valores absurdos.

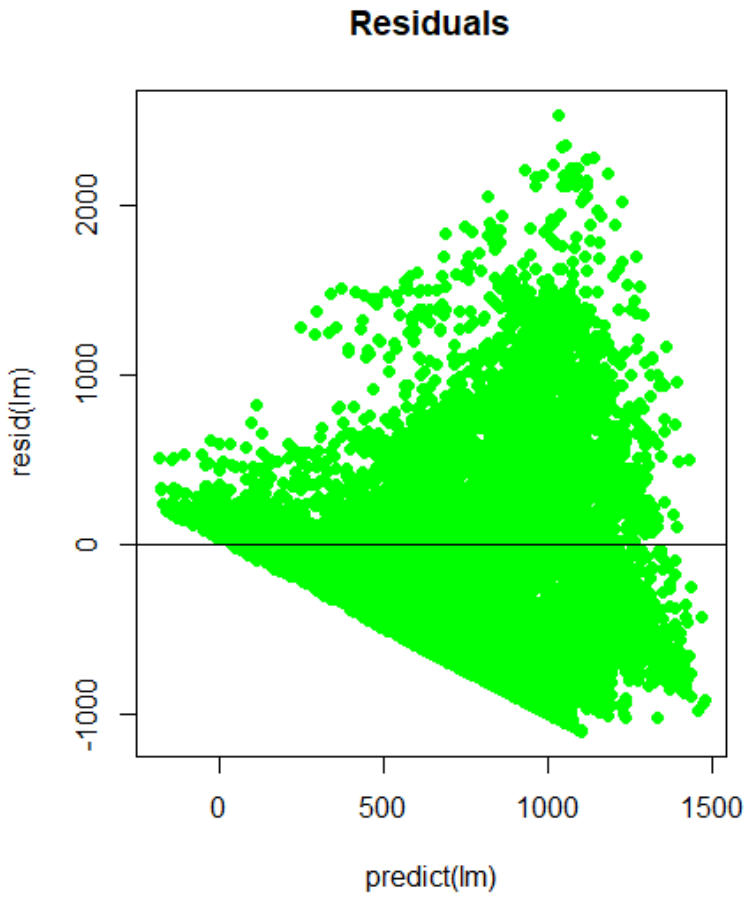

Figura 4 - Gráfico de resíduos do conjunto de dados inteiro

Regressão Linear Simples – Conjunto de dados separado por estações do ano

Como mencionado anteriormente, foi-nos requisitado a separação dos dados com recurso a uma variável qualitativa presente no nosso conjunto de dados, sendo que a nossa decisão, foi de optarmos pela variável denominada no conjunto de dados como "*Seasons*", que indica as estações do ano. Nos subtópicos respetivos, será efetuada a descrição da análise efetuada a cada subconjunto de dados.

Inverno

Iniciamos a análise deste subconjunto de dados relativo apenas ao Inverno, com a demonstração do diagrama de dispersão que nos pode indicar a existência de alguma correlação linear. Esse diagrama de dispersão, encontra-se presente na Figura 5 – Diagrama de Dispersão do subconjunto de dados (Inverno). Conseguimos visualizar efetivamente a existência de uma correlação linear positiva, mas como referido anteriormente, temos que utilizar mais métodos para comprovarmos a eventual existência dessa correlação.

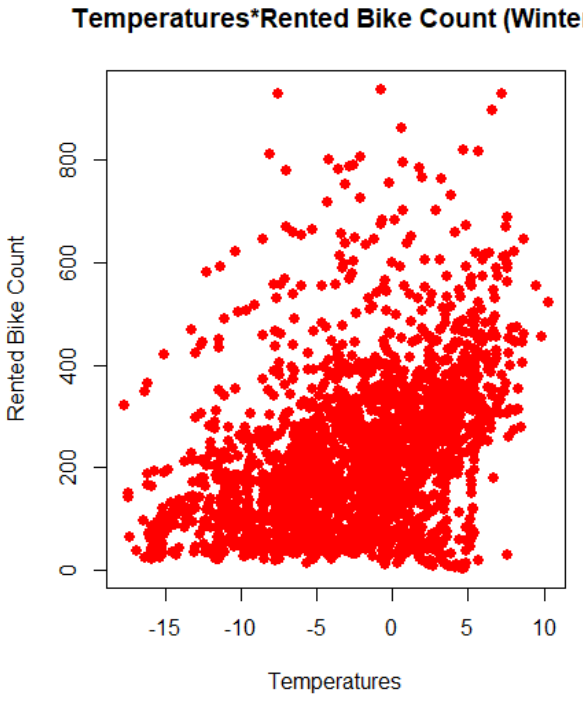

Figura 5 - Diagrama de Dispersão do subconjunto de dados (Inverno)

Como tal, recorremos de forma igual ao que fora retratado no conjunto de dados inteiro, ou seja, complementamos a análise gráfica, com o coeficiente de correlação linear de Pearson.

Após a execução do comando em R que indica o resultado da fórmula, obtivemos o seguinte valor para o coeficiente de correlação linear:

$$r_{xy} = 0.3828507$$

Que efetivamente, comprova a existência de uma correlação linear positiva, e segundo a tabela indicativa de intensidade de correlação, apresentada na Figura 2, pode-se considerar que existe uma correlação linear positiva fraca, mas adequadas, pois está dentro dos parâmetros referidos anteriormente.

Mas ainda nos é requisitado mais métodos para chegarmos à conclusão de que o modelo ajustado é adequado. Para tal necessitamos de calcular a reta de regressão e os resíduos.

Relativamente à reta de regressão, com recurso a linhas de código em R, chegamos à seguinte reta de regressão.

$$\hat{y} = 252.28 + 10.53x$$

É possível, visualizar agora, uma reta de regressão diagrama de dispersão, como pode ser verificado na Figura 6 – Diagrama de Dispersão do subconjunto de dados (Inverno) com reta de regressão.

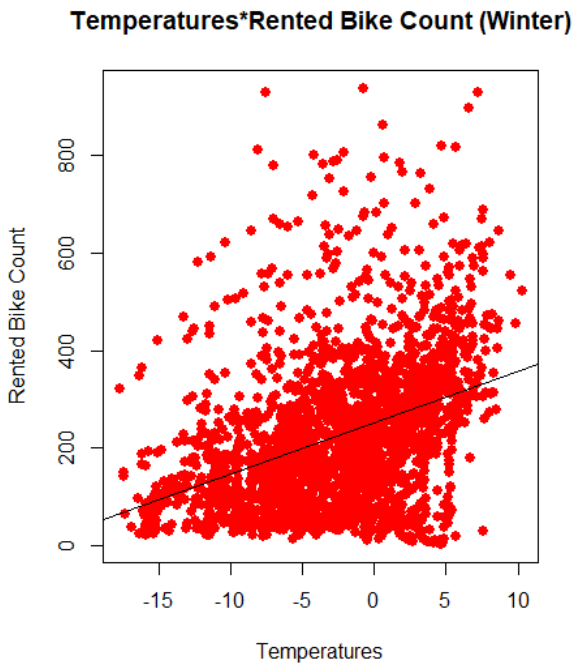

Figura 6 - Diagrama de Dispersão do subconjunto de dados (Inverno) com reta de regressão

Com essa reta de regressão podemos efetuar previsões relativamente a quantas bicicletas serão alugadas nas mais variadas temperaturas. Sendo que podemos por exemplo verificar para -10℃, que é um valor muito frequente no Inverno da cidade de Seoul:

$$\hat{\boldsymbol{y}} = \mathbf{252.28} + \mathbf{10.53} \times -\mathbf{10} = \mathbf{146.98}$$

Que é um valor que se encontra dentro do intervalo de alugueis de bicicletas. Entretanto, existem algumas temperaturas cujo valor de alugueis de bicicletas sejam negativos, sendo, portanto, um valor absurdo. Podemos indicar esse valor com por exemplo -25℃, que pertence ao intervalo de valores de temperatura que ocorrem no Inverno de Seoul:

$$\hat{\boldsymbol{y}} = \mathbf{252.28} + \mathbf{10.53} \times -\mathbf{25} = -\mathbf{10.97}$$

Também podemos de igual forma, efetuar previsões de temperatura através do número de bicicletas alugadas numa determinada data. Para tal, bastou-nos trocar os parâmetros no comando R que utilizamos para obter a correlação linear, e tal troca, deu-nos o seguinte resultado:

$$\hat{\boldsymbol{x}} = -5.68138 + 0.01393y$$

Com esta reta de regressão, conseguimos determinar algumas previsões para qual seria a temperatura com um determinado número de alugueis de bicicletas. Por exemplo, se existissem 500 alugueis, a temperatura seria:

$$\hat{\boldsymbol{x}}\,(500) = -5.68138 + 0.01393 \times 500 = 6.965$$

Ou para um número de alugueis mais frequente no Inverno, 100 alugueis:

$$\hat{\boldsymbol{x}}\,(100) = -5.68138 + 0.01393 \times 100 = -4.28838$$

Existem valores que acabam por ocorrer num valor absurdo, como por exemplo, quando existem 3500 alugueis de bicicletas (valor que pertence ao intervalo dos valores obtidos):

$$\hat{x}\,(3500) \;=\; -5.68138 \;+\; 0.01393 \;\times\; 3500 \;=\; 48.755$$

Sendo que este valor, é claramente absurdo, pois não está dentro dos valores de temperatura obtidos no trabalho anterior.

Relativamente aos resíduos, obtivemos o gráfico residual presente na Figura 7 – Gráfico de resíduos do subconjunto de dados (Inverno), que indica que não existe um padrão bem definido, portanto o modelo ajustado parece ser adequado. Embora o modelo tenha sido considerado adequado (diagrama de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados, criando assim uma previsão com valores absurdos.

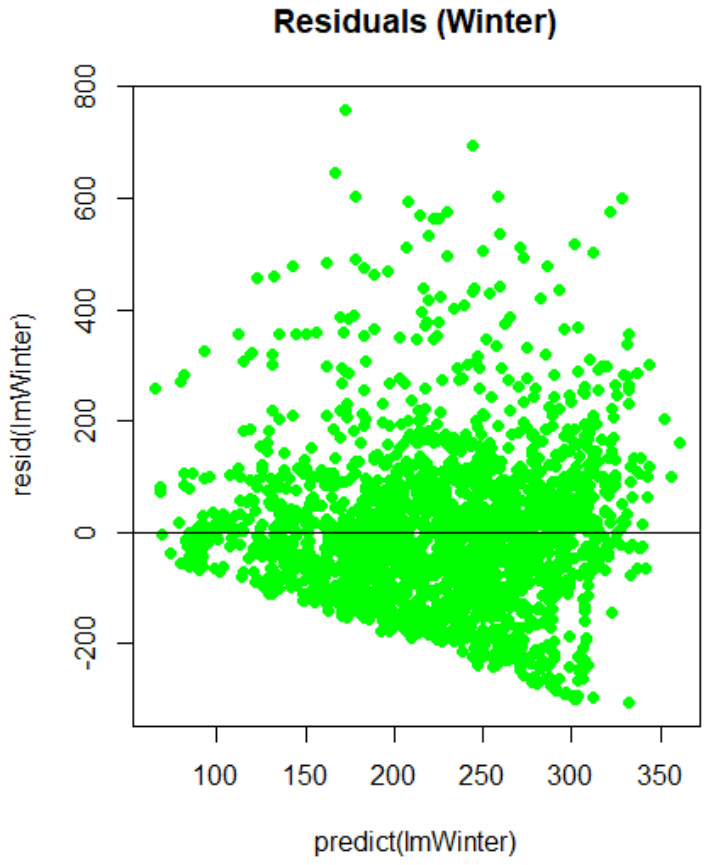

Figura 7 - Gráfico de resíduos do subconjunto de dados (Inverno)

Primavera

Iniciamos a análise deste subconjunto de dados relativo apenas à Primavera, com a demonstração do diagrama de dispersão que nos pode indicar a existência de alguma correlação linear. Esse diagrama de dispersão, encontra-se presente na Figura 8 – Diagrama de Dispersão do subconjunto de dados (Primavera). Conseguimos visualizar efetivamente a existência de uma correlação linear positiva, mas como referido anteriormente, temos que utilizar mais métodos para comprovarmos a eventual existência dessa correlação.

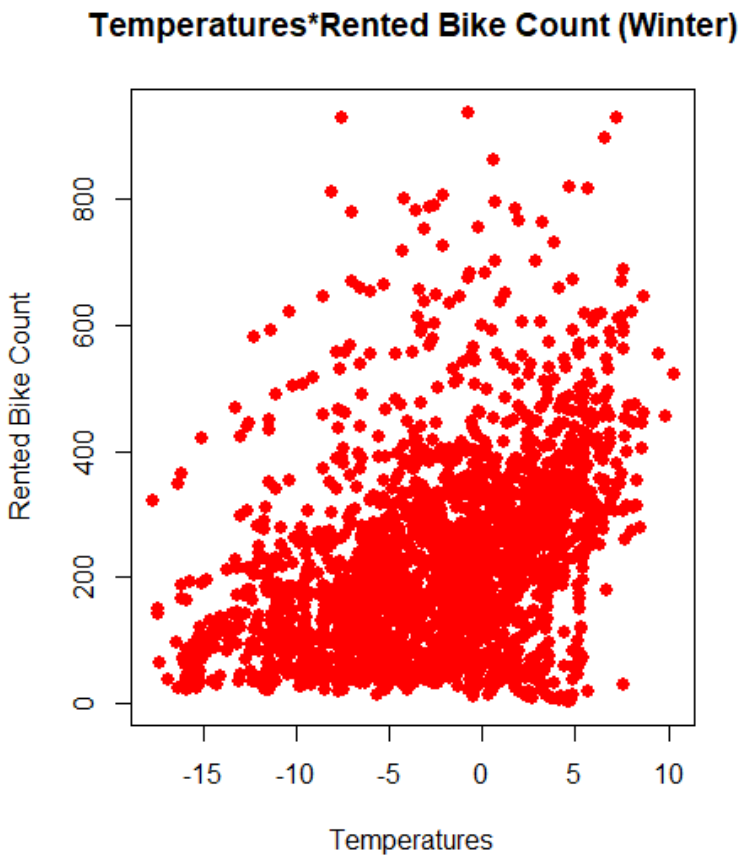

Figura 8 - Diagrama de Dispersão do subconjunto de dados (Primavera)

Como tal, recorremos de forma igual ao que fora retratado no conjunto de dados inteiro, ou seja, complementamos a análise gráfica, com o coeficiente de correlação linear de Pearson.

Após a execução do comando em R que indica o resultado da fórmula, obtivemos o seguinte valor para o coeficiente de correlação linear:

$$r_{xy} = 0.5582815$$

Que efetivamente, comprova a existência de uma correlação linear positiva, e segundo a tabela indicativa de intensidade de correlação, apresentada na Figura 2, pode-se considerar que existe uma correlação linear positiva moderada, mas adequadas, pois está dentro dos parâmetros referidos anteriormente.

Mas ainda nos é requisitado mais métodos para chegarmos à conclusão de que o modelo ajustado é adequado. Para tal necessitamos de calcular a reta de regressão e os resíduos.

Relativamente à reta de regressão, com recurso a linhas de código em R, chegamos à seguinte reta de regressão.

$$\hat{y} = 45.91 + 52.44x$$

É possível, visualizar agora, uma reta de regressão diagrama de dispersão, como pode ser verificado na Figura 9 – Diagrama de Dispersão do subconjunto de dados (Primavera) com reta de regressão.

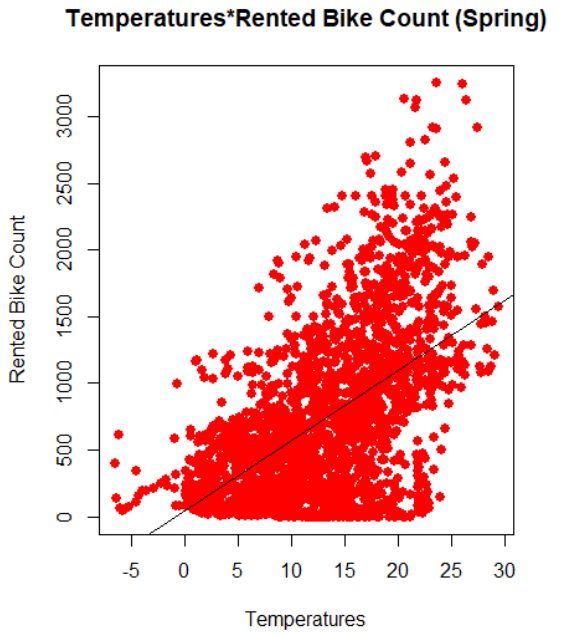

Figura 9 - Diagrama de Dispersão do subconjunto de dados (Primavera) com reta de regressão

Com essa reta de regressão podemos efetuar previsões relativamente a quantas bicicletas serão alugadas nas mais variadas temperaturas. Sendo que podemos por exemplo verificar para 10°C, que é um valor muito frequente na Primavera da cidade de Seoul:

$$\hat{\boldsymbol{y}} = \mathbf{45.91} + \mathbf{52.44} \times \mathbf{10} = 570.31$$

Que é um valor que se encontra dentro do intervalo de alugueis de bicicletas. Entretanto, existem algumas temperaturas cujo valor de alugueis de bicicletas sejam negativos, sendo, portanto, um valor absurdo. Podemos indicar esse valor com por exemplo -1°C, que pertence ao intervalo de valores de temperatura que ocorrem na Primavera de Seoul:

$$\hat{\boldsymbol{y}} = \mathbf{45.91} + \mathbf{52.44} \times -\mathbf{1} = -6.53$$

Também podemos de igual forma, efetuar previsões de temperatura através do número de bicicletas alugadas numa determinada data. Para tal, bastou-nos trocar os parâmetros no comando R que utilizamos para obter a correlação linear, e tal troca, deu-nos o seguinte resultado:

$$\hat{\boldsymbol{x}} = 8.707410 + 0.005944y$$

Com esta reta de regressão, conseguimos determinar algumas previsões para qual seria a temperatura com um determinado número de alugueis de bicicletas. Por exemplo, se existissem 500 alugueis, a temperatura seria:

$$\hat{\boldsymbol{x}}\,(500) = 8.707410 + 0.005944 \times 500 = 11.67941$$

Ou para um número de alugueis mais frequente na Primavera, 100 alugueis:

$$\hat{\boldsymbol{x}}\,(100) = 8.707410 + 0.005944 \times 100 = 9.30181$$

Relativamente aos resíduos, obtivemos o gráfico residual presente na Figura 10 – Gráfico de resíduos do subconjunto de dados (Primavera), que indica que não existe um padrão bem definido, portanto o modelo ajustado parece ser adequado. Embora o modelo tenha sido considerado adequado (diagrama de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados, criando assim uma previsão com valores absurdos.

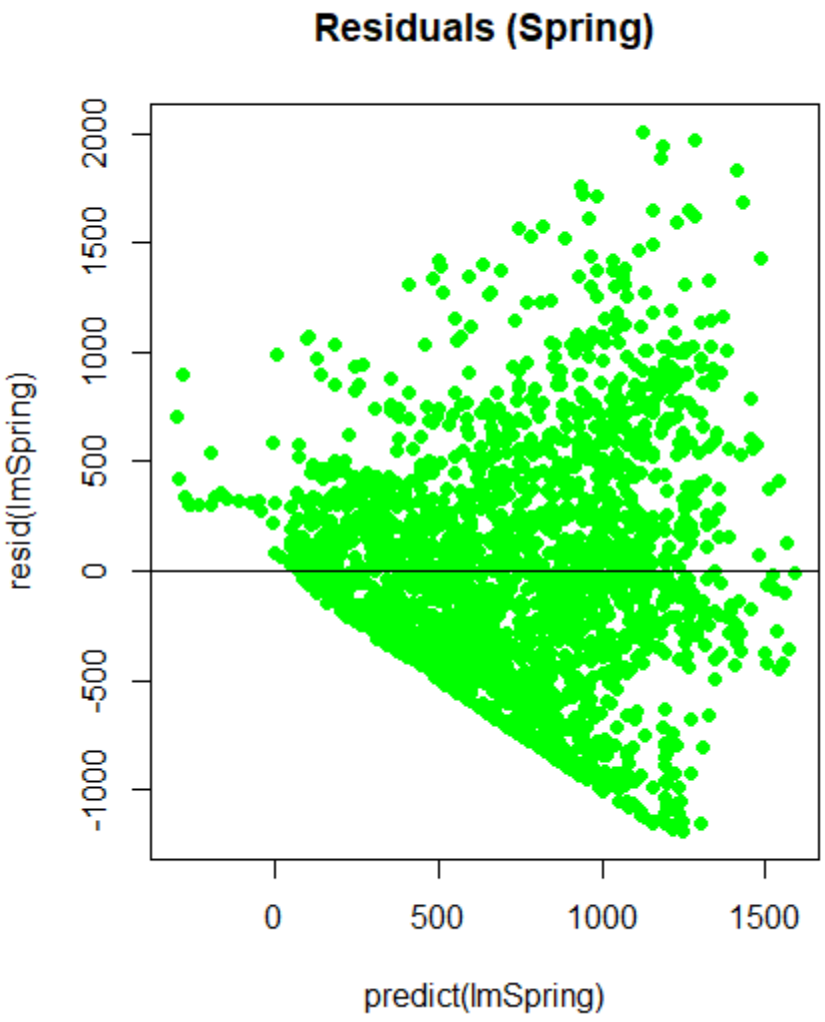

Figura 10 - Gráfico de resíduos do subconjunto de dados (Primavera)

Outono

Iniciamos a análise deste subconjunto de dados relativo apenas ao Outono, com a demonstração do diagrama de dispersão que nos pode indicar a existência de alguma correlação linear. Esse diagrama de dispersão, encontra-se presente na Figura 11 – Diagrama de Dispersão do subconjunto de dados (Outono). Conseguimos visualizar efetivamente a existência de uma correlação linear positiva, mas como referido anteriormente, temos que utilizar mais métodos para comprovarmos a eventual existência dessa correlação.

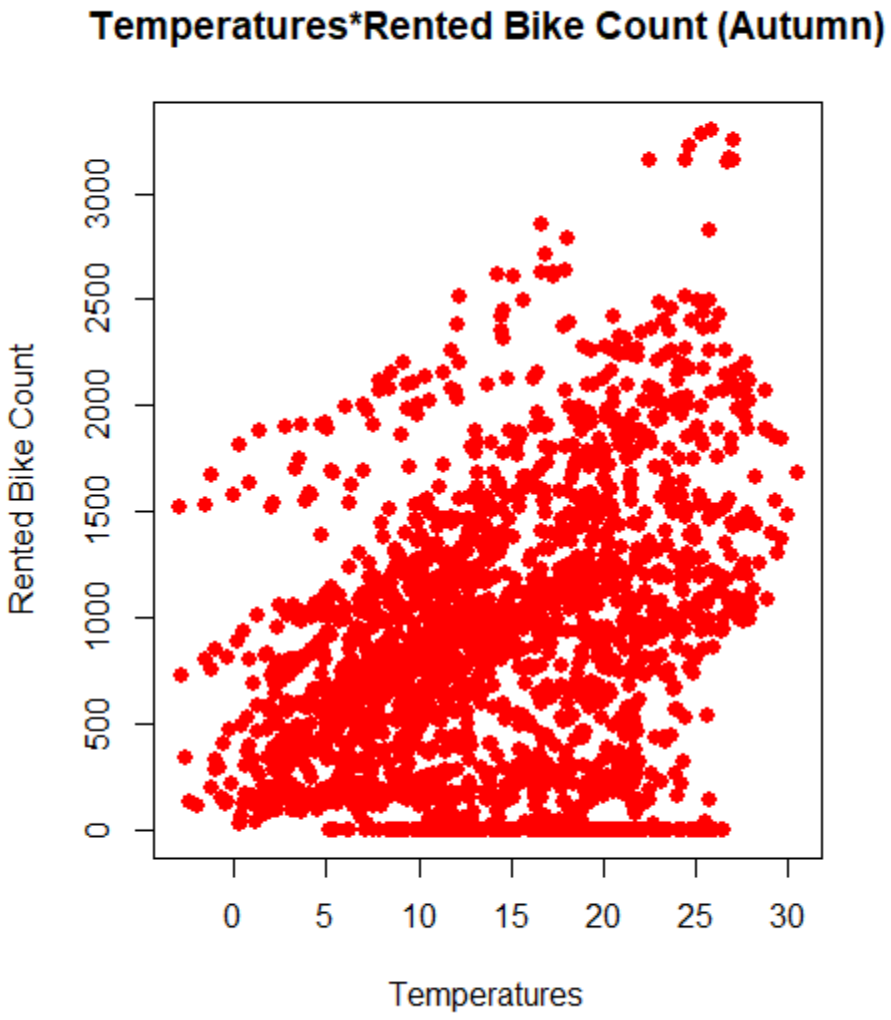

Figura 11 - Diagrama de Dispersão do subconjunto de dados (Outono)

Como tal, recorremos de forma igual ao que fora retratado no conjunto de dados inteiro, ou seja, complementamos a análise gráfica, com o coeficiente de correlação linear de Pearson.

Após a execução do comando em R que indica o resultado da fórmula, obtivemos o seguinte valor para o coeficiente de correlação linear:

$$r_{xy} = 0.3177157$$

Que efetivamente, comprova a existência de uma correlação linear positiva, e segundo a tabela indicativa de intensidade de correlação, apresentada na Figura 2, pode-se considerar que existe uma correlação linear positiva fraca, mas adequadas, pois está dentro dos parâmetros referidos anteriormente.

Mas ainda nos é requisitado mais métodos para chegarmos à conclusão de que o modelo ajustado é adequado. Para tal necessitamos de calcular a reta de regressão e os resíduos.

Relativamente à reta de regressão, com recurso a linhas de código em R, chegamos à seguinte reta de regressão.

$$\hat{y} = 405.99 + 29.29x$$

É possível, visualizar agora, uma reta de regressão diagrama de dispersão, como pode ser verificado na Figura 12 – Diagrama de Dispersão do subconjunto de dados (Outono) com reta de regressão.

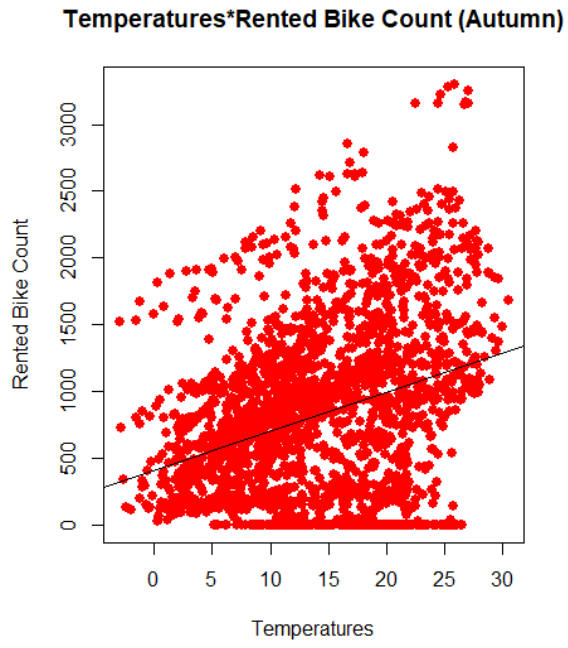

Figura 12 - Diagrama de Dispersão do subconjunto de dados (Outono) com reta de regressão

Com essa reta de regressão podemos efetuar previsões relativamente a quantas bicicletas serão alugadas nas mais variadas temperaturas. Sendo que podemos por exemplo verificar para 10℃, que é um valor muito frequente no Outono da cidade de Seoul:

$$\hat{\boldsymbol{y}} = 405.99 + 29.29 \times \mathbf{10} = 698.89$$

Que é um valor que se encontra dentro do intervalo de alugueis de bicicletas. Entretanto, existem algumas temperaturas cujo valor de alugueis de bicicletas sejam negativos, sendo, portanto, um valor absurdo. Podemos indicar esse valor com por exemplo -14℃, que pertence ao intervalo de valores de temperatura que ocorrem no Outono de Seoul:

$$\hat{\boldsymbol{y}} = 405.99 + 29.29 \times -\mathbf{14} = -4.07$$

Também podemos de igual forma, efetuar previsões de temperatura através do número de bicicletas alugadas numa determinada data. Para tal, bastou-nos trocar os parâmetros no comando R que utilizamos para obter a correlação linear, e tal troca, deu-nos o seguinte resultado:

$$\hat{\boldsymbol{x}} = 11.296295 + 0.003446y$$

Com esta reta de regressão, conseguimos determinar algumas previsões para qual seria a temperatura com um determinado número de alugueis de bicicletas. Por exemplo, se existissem 500 alugueis, a temperatura seria:

$$\hat{\boldsymbol{x}}\,(500) = 11.296295 + 0.003446 \times 500 = 13.019295$$

Ou para um número de alugueis mais frequente no Outono, 100 alugueis:

$$\hat{\boldsymbol{x}}\,(100) = 8.707410 + 0.005944 \times 100 = 9.30181$$

Relativamente aos resíduos, obtivemos o gráfico residual presente na Figura 13 – Gráfico de resíduos do subconjunto de dados (Outono), que indica que não existe um padrão bem definido, portanto o modelo ajustado parece ser adequado. Embora o modelo tenha sido considerado adequado (diagrama de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados, criando assim uma previsão com valores absurdos.

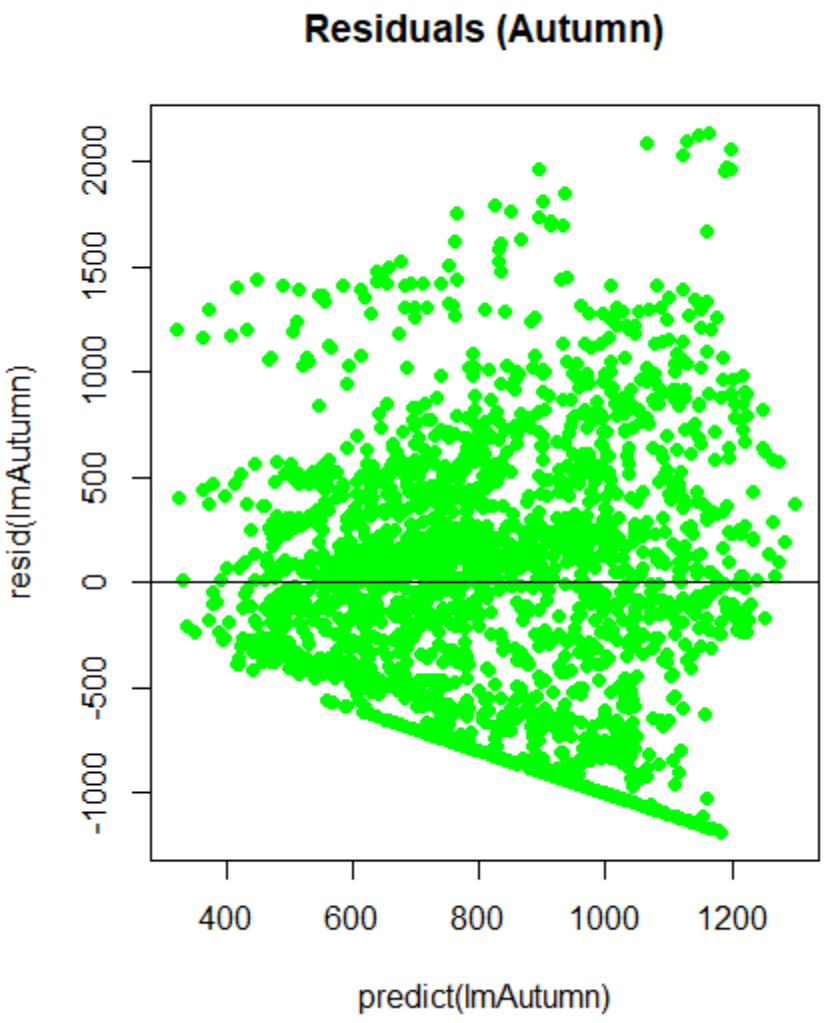

Figura 13 - Gráfico de resíduos do subconjunto de dados (Outono)

Verão

Iniciamos a análise deste subconjunto de dados relativo apenas ao Verão, com a demonstração do diagrama de dispersão que nos pode indicar a existência de alguma correlação linear. Esse diagrama de dispersão, encontra-se presente na Figura 14 – Diagrama de Dispersão do subconjunto de dados (Verão). Conseguimos visualizar efetivamente a existência de uma correlação linear positiva, mas como referido anteriormente, temos que utilizar mais métodos para comprovarmos a eventual existência dessa correlação.

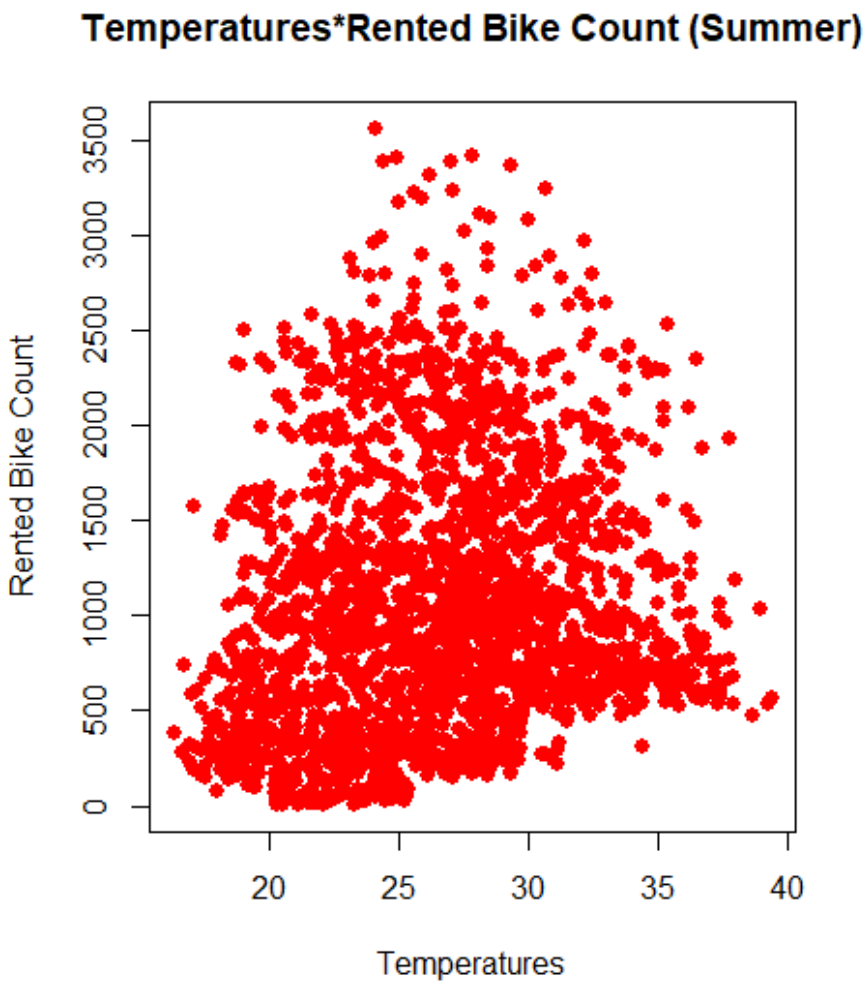

Figura 14 - Diagrama de Dispersão do subconjunto de dados (Verão)

Como tal, recorremos de forma igual ao que fora retratado no conjunto de dados inteiro, ou seja, complementamos a análise gráfica, com o coeficiente de correlação linear de Pearson.

Após a execução do comando em R que indica o resultado da fórmula, obtivemos o seguinte valor para o coeficiente de correlação linear:

$$r_{xy} = 0.1634651$$

Que efetivamente, comprova a existência de uma correlação linear positiva, e segundo a tabela indicativa de intensidade de correlação, apresentada na Figura 2, pode-se considerar que existe uma correlação linear positiva fraca, mas adequadas, pois está dentro dos parâmetros referidos anteriormente.

Mas ainda nos é requisitado mais métodos para chegarmos à conclusão de que o modelo ajustado é adequado. Para tal necessitamos de calcular a reta de regressão e os resíduos.

Relativamente à reta de regressão, com recurso a linhas de código em R, chegamos à seguinte reta de regressão.

$$\hat{y} = 390.8 + 24.2x$$

É possível, visualizar agora, uma reta de regressão diagrama de dispersão, como pode ser verificado na Figura 15 – Diagrama de Dispersão do subconjunto de dados (Verão) com reta de regressão.

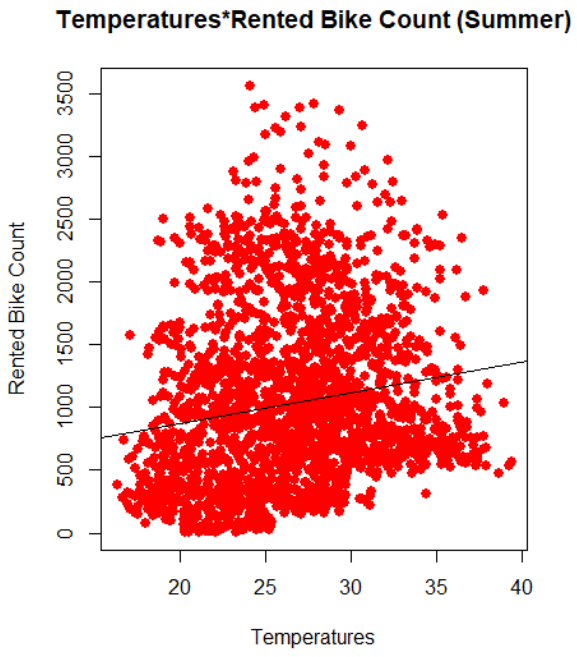

Figura 15 - Diagrama de Dispersão do subconjunto de dados (Verão) com reta de regressão

Com essa reta de regressão podemos efetuar previsões relativamente a quantas bicicletas serão alugadas nas mais variadas temperaturas. Sendo que podemos por exemplo verificar para 20ºC, que é um valor muito frequente no Verão da cidade de Seoul:

$$\widehat{\boldsymbol{y}} = 390.8 + 24.2 \times \mathbf{20} = 874.8$$

Também podemos de igual forma, efetuar previsões de temperatura através do número de bicicletas alugadas numa determinada data. Para tal, bastou-nos trocar os parâmetros no comando R que utilizamos para obter a correlação linear, e tal troca, deu-nos o seguinte resultado:

$$\widehat{\boldsymbol{x}} = 25.441033 + 0.001104y$$

Com esta reta de regressão, conseguimos determinar algumas previsões para qual seria a temperatura com um determinado número de alugueis de bicicletas. Por exemplo, se existissem 500 alugueis, a temperatura seria:

$$\widehat{\boldsymbol{x}}(500) = 25.441033 + 0.001104 \times 500 = 25.993033$$

Ou para um número de alugueis mais frequente no Verão, 2500 alugueis:

$$\widehat{\boldsymbol{x}}(100) = 25.441033 + 0.001104 \times 2500 = 28.201033$$

Relativamente aos resíduos, obtivemos o gráfico residual presente na Figura 16 – Gráfico de resíduos do subconjunto de dados (Verão), que indica que não existe um padrão bem definido, portanto o modelo ajustado parece ser adequado. Embora o modelo tenha sido considerado adequado (diagrama de dispersão, coeficiente de correlação linear e resíduos, todos levaram a considerar o modelo adequado), nada nos garante que a reta obtida se mantém para valores afastado dos observados, criando assim uma previsão com valores absurdos.

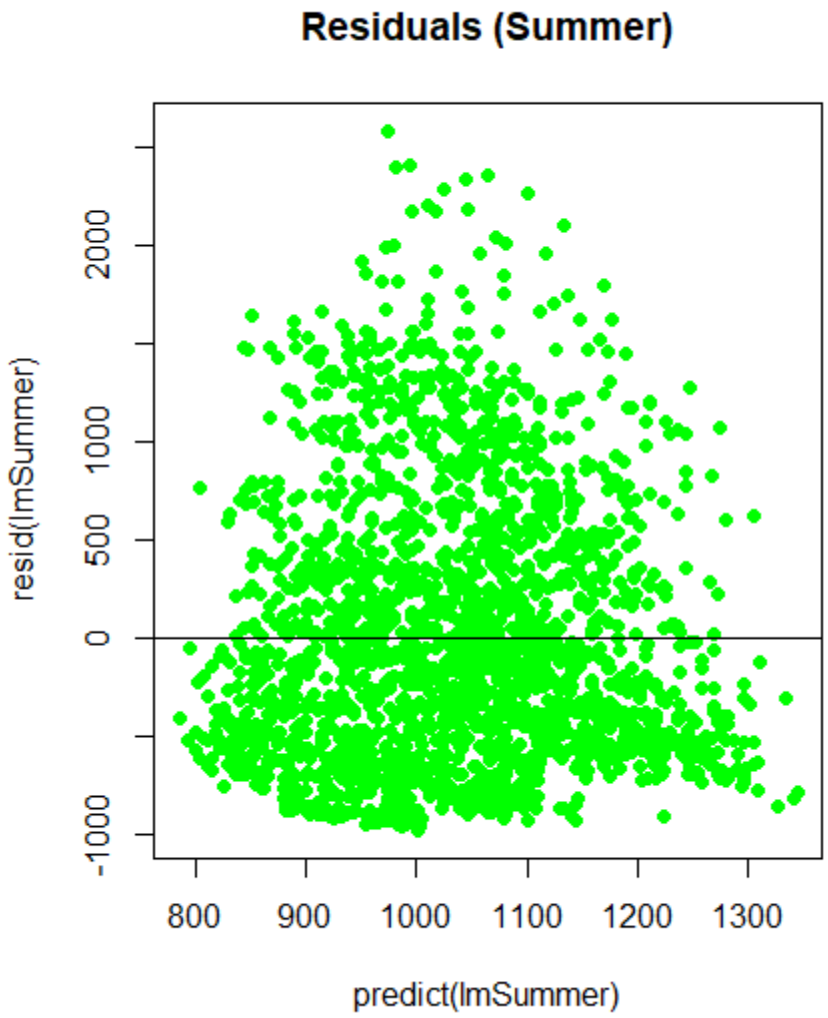

Figura 16 - Gráfico de resíduos do subconjunto de dados (Verão)

Conclusão

Com esta análise, conseguimos determinar a existência de uma variação entre os modelos ajustados, e conseguimos também verificar a existência de uma correlação positiva nas mais variadas estações do ano. Cremos como grupo, que as escolhas da variável independente e da variável dependente tenham uma relevância na adequação dos modelos, e na prova da existência de uma relação entre as duas variáveis escolhidas. Também acreditamos que, a existência de uma intensidade de correlação inferior em algumas estações do ano possa ter a ver com o facto de ser as estações do ano mais irregulares quanto a temperaturas, e que não permitem termos um conjunto de dados minimamente fiável para ajustamento de dados.

A aprendizagem que retiramos desta análise sucinta, dar-nos-á uma maior valência, caso surja interesse de algum dos membros do grupo de seguir a área de probabilidade e estatística profissionalmente, como também certamente, dar-nos-á um incremento no leque de capacidades que adquirimos na trajetória do curso.

Referências bibliográficas

Bibliografia/Netgrafia

- Slides disponibilizados pelos professores, no moodle (Capítulo 2 – Regressão Linear Simples).
- Fichas das aulas práticas.