UNIVERSIDADE FEDERAL DO PARANÁ

Rafael Morciani Alves da Silva

Lais Hoffman

Relatório Regressão - Regressão linear múltipla com dados de precipitação da cidade de Morretes-PR

CURITIBA

12 de Junho de 2018

Resumo

Temos como objetivo aplicar Regressão Linear múltipla, sendo nossa variável resposta a precipitação (chuva). Com nosso escopo definido conseguimos partir para analise exploratória de todas as variáveis disponíveis e viáveis do nosso banco de dados. Assim vamos utilizar os métodos como inferência, testes para verificar os pressupostos exigidos pelo tipo de modelo usado, obtendo quais variáveis que são realmente significativas agregando valor ao modelo ou até mesmo analisar as possíveis multicolinearidade que é possível observar na correlação das mesmas, pois isso pode afetar de maneira negativa o modelo. Passando por vários testes e registrando evidências logo todos os passos devem ter sempre uma justificativa plausível e coerente para que a tomada de decisões seja de fato eficaz, assim desta forma chegando no modelo mais adequado.

Introdução

Regressão linear múltipla é um conjunto de métodos estatísticos para construção de modelos que descrevem *parcialmente* a relação entre uma variável resposta com duas ou mais variáveis explicativas. Na regressão linear simples a única diferença é que a variável resposta é explicada somente por uma variável explicativa.

Neste estudo será utilizado a regressão linear múltipla, que terá como objetivo explicar a precipitação (chuva) em relação às demais variáveis explicativas da base de dados utilizada.

Materiais e Métodos

Materiais

A base de dados analisada foi adquirida no site (http://www.inmet.gov.br/portal/index.php?r=estacoes/estacoesAutomaticas). A análise será realizada nos dados da cidade de Morretes-PR que foram coletados entre os dias 22/05/2017 e 22/05/2018, a base contém observações de hora em hora.

Foi utilizado também como suporte ao estudo os materiais apresentados em sala de aula pelo professor Cesar Taconeli que estão disponíveis em (https://docs.ufpr.br/~taconeli/CE07118/CE07118.html).

Para aplicação dos métodos se fez necessária a utilização do software **R**, onde nele ainda foram utilizados dois pacotes adicionais, o pacote *car* e o pacote *faraway*.

Métodos

O modelo de regressão linear múltipla é definido por:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \epsilon_i. \tag{1}$$

Onde:

y_i: Valor da variável resposta observada na i-ésima observação.

β_0: Intercépto do modelo (média de y quando as demais variáveis são iguais a zero).

$\beta_i,\ 0 < i < k$: Alteração esperada em y para uma unidade a mais em x_i, quando as demais covariáveis são fixas.

ϵ_i: Erro aleatório associado a i-ésima observação.

Para a utilização da análise de regressão linear múltipla precisamos assumir alguns pressupostos sobre o erro a priori:

- $\mathrm{E}(\epsilon_i) = 0$
- $\mathrm{Var}(\epsilon_i) = \sigma^2$
- $\epsilon_i \perp \epsilon_j, i \neq j$
- $x_i \perp \epsilon_i, \forall i$
- $\epsilon_i \sim N(0; \sigma^2)$

Importação e organização da base de dados:

Após a base de dados ser importada no R, foi necessário organizá-la para melhor aplicação dos métodos, para isso as seguntes mudanças foram feitas na base de dados:

- Foram omitidas todas as linhas que possúiam observações faltando (NA's).
- OS dados foram agrupados pela média do dia, com a função *aggregate* do R.
- Foram excluidas as linhas onde não foi observada nenhuma chuva no dia, para não acontecer da base ser inflacionada de zeros dificultado a aplicação do método.
- Algumas variáveis foram deixadas de lado, entre elas: Variáveis com valores de máximos e mínimos observados no dia, variáveis categóricas (data e código da estação), hora (variáveis agrupadas por dia) e humidade (por ser ambiguo com a variável resposta precipitação).

Construção do modelo:

O 1º modelo foi ajustado com todas as covariáveis da base, e os respectivos $\beta's$ foram estimados com a função *lm* do R, resultando no seguinte modelo de regressão linear múltipla:

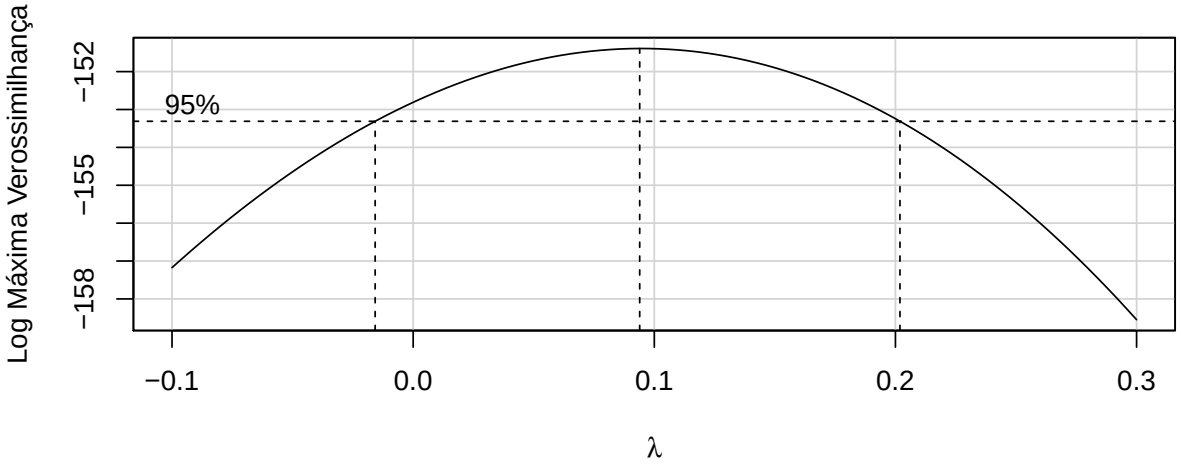
- y = -61.12 -0.50∗temperatura +0.81∗ponto de orvalho +0.054∗pressão -4.57∗direção do vento -0.00025∗velocidade do vento +2.34 ∗rajada de vento -0.00075∗radiacao + ϵ_i

Porém, pelo *summary* do modelo 1, foi observado que algumas covariáveis não são significativas, para isso um 2º modelo foi gerado, desta vez utilizando a função *step* do R, pelo método stepwise.

2º modelo ajustado:

- y = -5.88 -0.53∗temperatura +0.76∗ponto de orvalho -4.64∗direção do vento +2.40∗rajada de vento + ϵ_i

Gerado o modelo e fixadas as variáveis explicativas, foram verificados os pressupostos de normalidade e homocedasticidade dos resíduos. O teste de normalidade foi realizado pelo teste de Shapiro-Wilk (*shapiro.test*), e o teste de homocedasticidade foi realizado pelo teste de pontuação (*ncvTest* do pacote car). Ambos testes deram resultados significativos a 5% (rejeitando-se H_0), indicando que os resíduos não se distribuem normalmente e nem possuem variância constante, para contornar o problema uma transformação do tipo $y' = y^\lambda$ na variável resposta foi necessária, onde λ é estimado por máxima verossimilhança pela função *boxCox* conforme mostra a figura abaixo:



O método de Box-Cox retornou valores de λ no intervado [-0.2 ; 0.2], como o valor zero está contido no intervalo, uma transformação do tipo $y' = log(y)$ foi realizada. Retornando o modelo final:

- y = -3.14 -0.26 $*$ temperatura +0.33 $*$ ponto de orvalho -0.66 $*$ direção do vento +0.59 $*$ rajada de vento + ϵ_i

Novamente verificados os pressupostos, desta vez nenhum significativo, indicando normalidade e heterocedasticidade dos resíduos do novo ajuste.

Foi também verificada a correlação entre as covariáveis, somente duas covariáveis apresentaram alto nível de correlação (temperatura e ponto de orvalho), essa correlação alta é devido ao fato de que o ponto de orvalho é a temperatura a qual o vapor de água que está em suspensão no ar começa a se condensar (virar "orvalho"). Segue abaixo a matriz de correlação de Pearson entre as variáveis explicativas do modelo final:

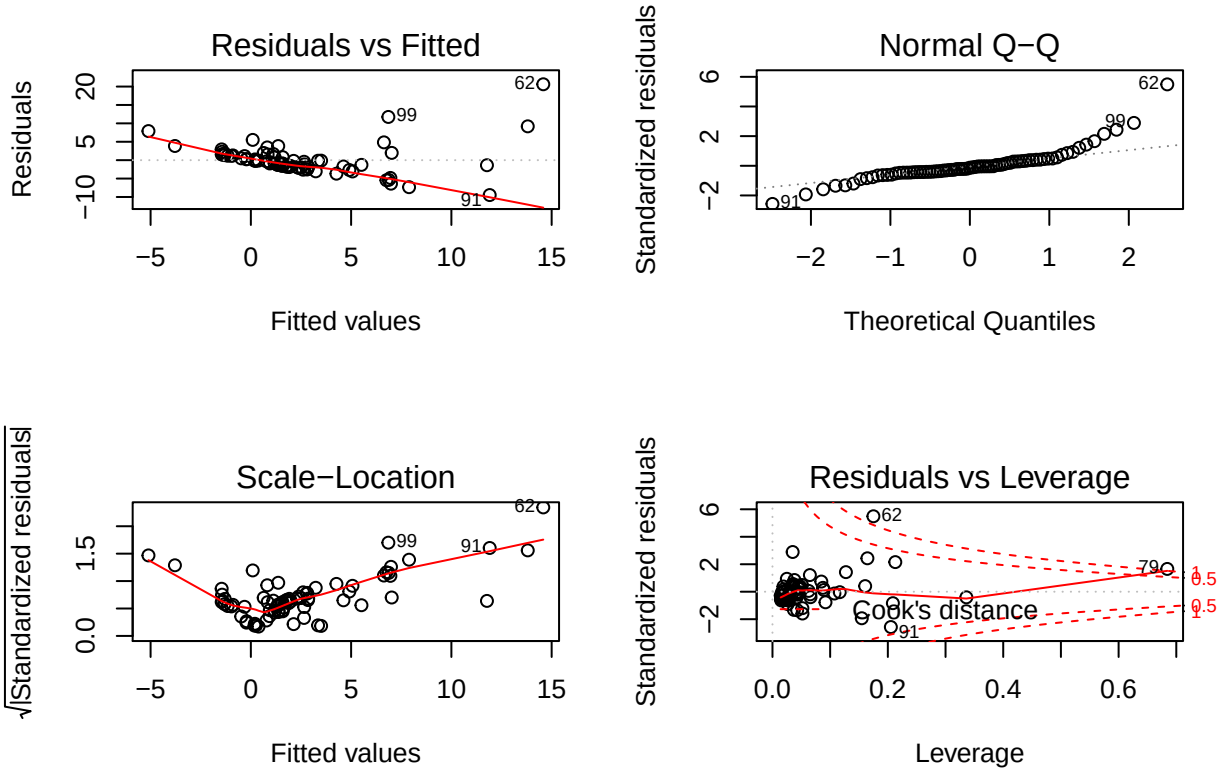
	Temperatura	Ponto orvalho	Direção vento	Velocidade vento
Temperatura	1.00000000	0.90253044	0.06091559	0.02804630
Ponto orvalho	0.90253044	1.00000000	-0.03054600	0.04499535
Direção vento	0.06091559	-0.03054600	1.00000000	-0.34135002
Velocidade vento	0.02804630	0.04499535	-0.34135002	1.00000000

Resultados e Discussões

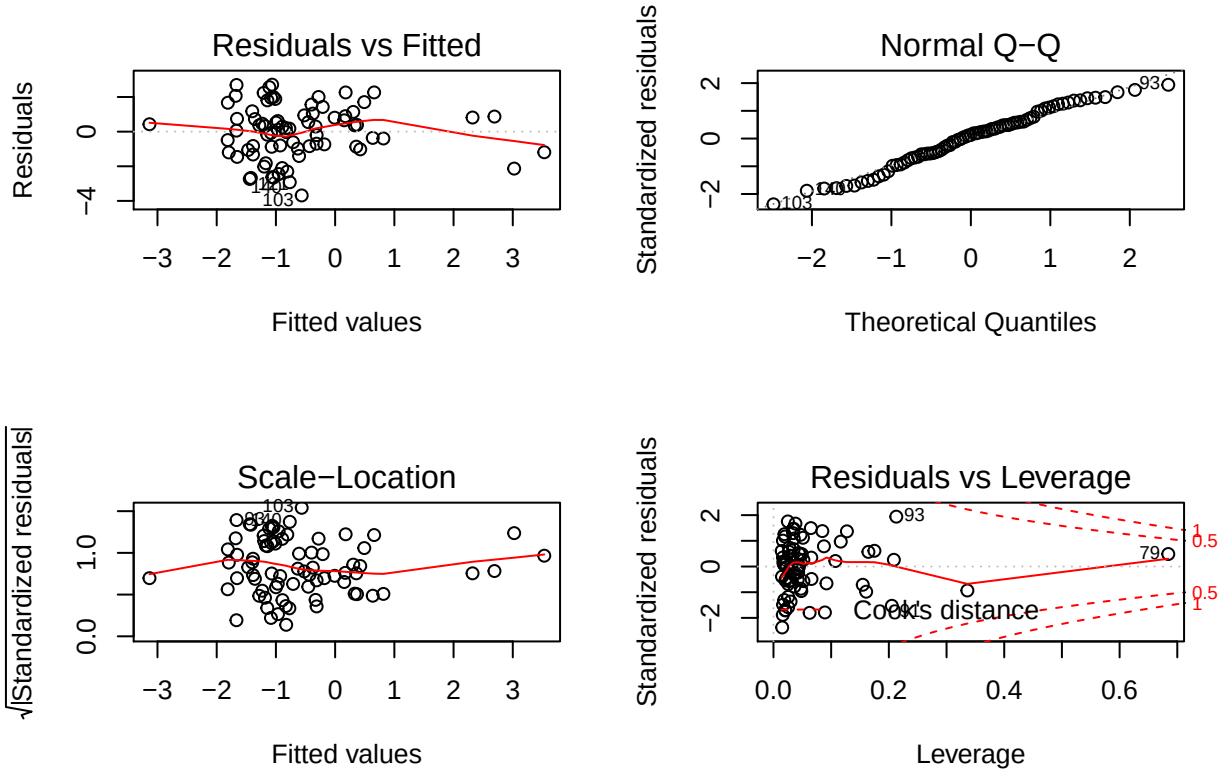
Neste estudo foram aplicadas as ferramentas e técnicas apresentadas em sala de aula na discilina de análise de regressão linear em um problema real, dentro dos limites de conhecimento foi possível ajustar um modelo de regressão linear múltipla para descrever o comportamento de uma variável resposta (precipitação), em relação a quatro variáveis explicativas (temperatura, Ponto de orvalho, Direção do vento e Velocidade do vento). O modelo ajustado atendeu os pressupostos necessários (normalidade e homocedasticidade) após uma transformação do tipo $y' = log(y)$ na variável resposta.

Comparando os ajustes:

Ajuste 2: Variável não transformada

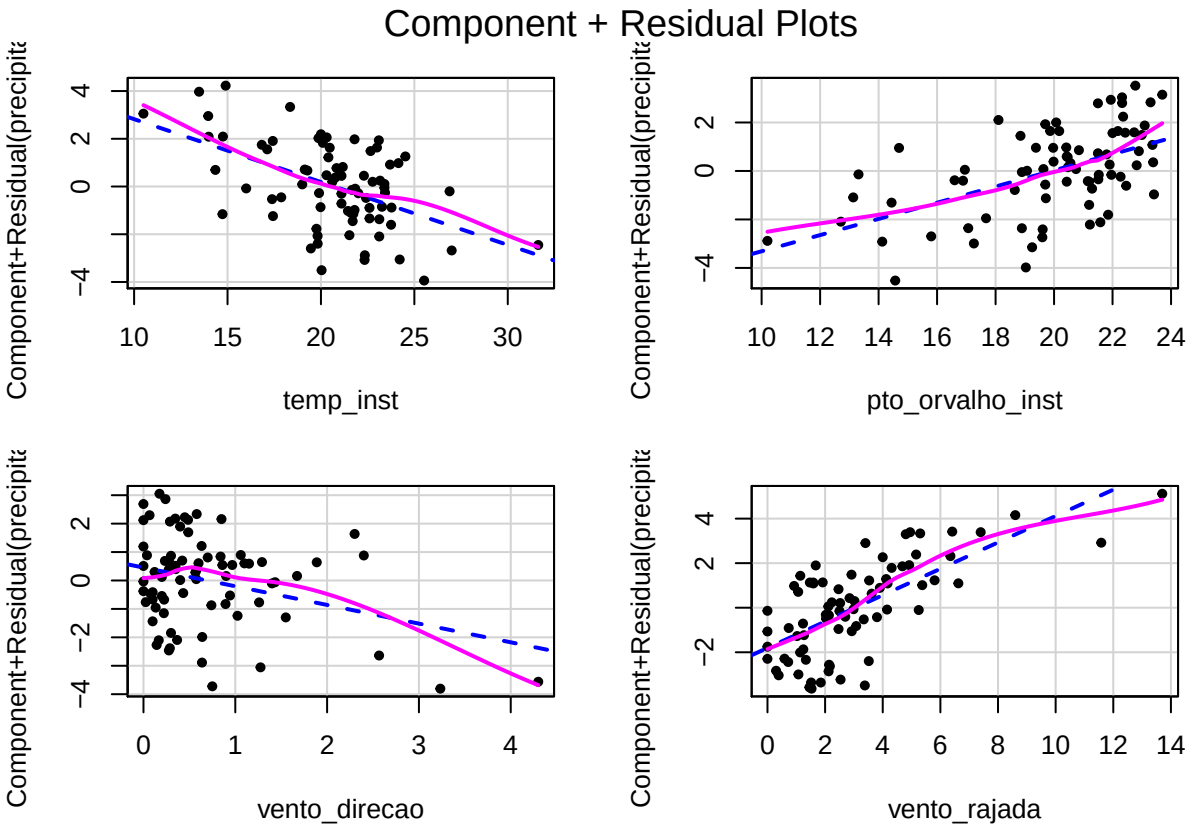


Ajuste 3: Variável transformada



Neste gráfico podemos notar facilmente a mudança entre o modelo 2 e 3, primeiramente a variabilidade dos resíduos, onde no modelo 2 é obvia a não homogeneidade da variância quanto no modelo 3 os resíduos estão mais dispersos e apresentam homogeneidade na variância. Quando aos gráficos de normalidade dos resíduos, no ajuste 2 os dados estão bem dispersos da linha que indica normalidade, e no ajuste 3 os dados se distribuem em cima da linha de normalidade. O mesmo acontece para os demais gráficos, os resíduos do modelo 3 tem comportamento desejádo, quanto no modelo 2 não.

Gráfico de resíduos parciais do modelo final:



Este gráfico mostra como os resíduos parciais de cada covariável se comportam em relação a variável resposta, é possível observar que somente a variável direção do vento possui um desajuste, este desajuste pode ser em função das observações muito extremas na variável explicativa, podendo gerar um ponto de alavanca.