Análise De regressão para o conjunto de dados AUTO

Samuel Martins de Medeiros

Introdução.

A análise de regressão múltipla é uma técnica estatística amplamente utilizada em diversas áreas, desde a economia até a biologia, para estudar a relação entre uma variável dependente e várias variáveis independentes. Neste trabalho, será aplicado a análise de regressão múltipla ao conjunto de dados "Auto" do pacote ISLR no R para investigar a relação entre a variável MPG (milhas por galão) e outras variáveis independentes, como a potência do motor, peso e aceleração.

Seguindo então para a análise exploratória do conjunto, será realizado uma análise descritiva dos dados juntamente a *plots* gráficos de dispersão para verificar a relação entre MPG e as variáveis independentes, bem como distribuição ou possíveis *outliers*. Também será se há valores ausentes no conjunto de dados.

Na seção de ajuste de modelo, a regressão múltipla com a variável MPG e as demais como variáveis independentes, bem como a seleção do subconjunto do total de variáveis que retorna o modelo mais parcimonioso. Avaliando, em seguida, a qualidade do ajuste do modelo usando medidas como p-valor e AIC.

Análise Exploratória.

A análise exploratória é considerada uma parte fundamental de qualquer tipo de análise dentro do âmbito da estatística e análise de dados. A identificação de padrões ou possíveis inconsistências pode ser vista durante a análise, bem como possíveis distribuições para os dados ou até mesmo erros que possam surgir durante as outras etapas da modelagem.

A primeira etapa pode ser considerada como a identificação da estrutura dos dados, bem como a presença de observações faltantes dentro do conjunto. Para o conjunto de dados "Auto", não foram identificados dados faltantes. É possível verificar a disposição dos dados por meio da Tabela 1, que apresenta as 10 primeiras observações como exemplo.

Table 1: Conjunto de dados Auto

mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
18	8	307	130	3504	12.0	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11.0	70	1	plymouth satellite
16	8	304	150	3433	12.0	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10.0	70	1	ford galaxie 500

É apresentado na estrutura um conjunto de 7 variáveis numéricas (mpg, cylinders, displacement, horsepower, weight, acceleration, year) e 2 variáveis do tipo fator (name, origin). A variável 'name', por apresentar um total de 304 valores únicos, sendo o conjunto de dados formado por 392 observações, será desconsiderada na análise. As variáveis seguem sendo:

Table 2: Variáveis e descrição

Variável	Descrição
mpg	Milhas por Galão.
cylinders	Número de cilindros, entre 4 e 8.
displacement	Deslocamento do motor.
horsepower	Potência do motor.
weight	Peso do veículo.
acceleration	Tempo de aceleração de 0 a 60mph(Segundos).
year	Ano do modelo.
origin	Origem do carro (1. Americano, 2. Europeu, 3. Japônes).

É possível identificar pelos gráficos de dispersão a presença de uma relação entre as variáveis explicativas e a variável dependente, note que até mesmo para variáveis inteiras ou fator, essa relação ainda existe, sendo mais acentuada para as variáveis *horsepower* e *weigth*, enquanto que na variável *acceleration* podemos identificar uma dispersão mais concisa para valores baixos que se dispersam mais conforme o valor de acceleration é aumentado. É possível ver, também, o tipo de relação, positiva para as variáveis *acceleration*, *year* e *origin*, e uma relação negativa para as demais.

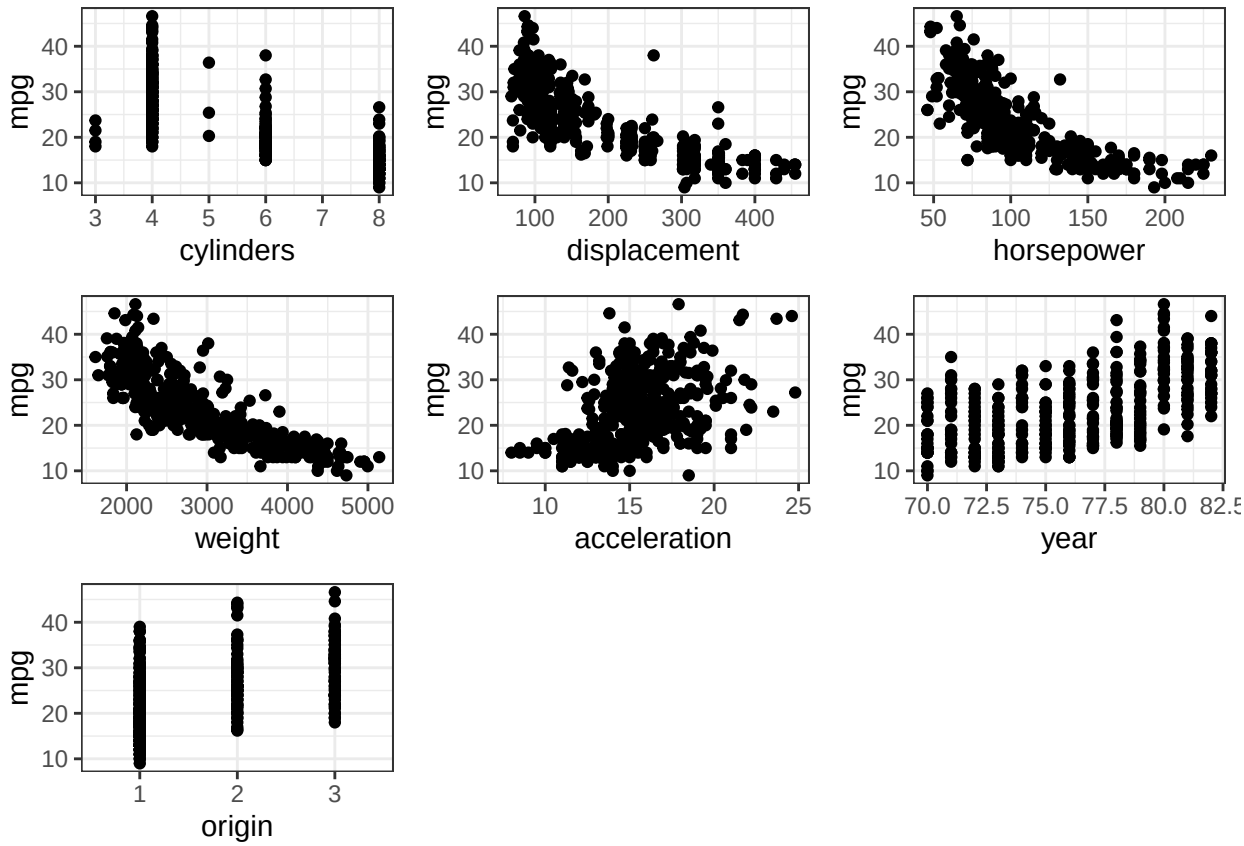


Figure 1: Gráficos de dispersão das covariáveis em relação a variável resposta

Esse fator da relação, positiva ou negativa, ou ainda a itensidade dessa relação pode ser identificada pela análise da correlação entre as variáveis, como pode ser visto as hipóteses antes citadas a partir dos gráficos seguem sendo verdadeiras pela análise da correlação das variáveis numéricas do conjunto de dados.

Table 3: Correlação entre as variáveis

	mpg	cylinders	displacement	horsepower	weight	acceleration	year
mpg	1.00	-0.78	-0.81	-0.78	-0.83	0.42	0.58
cylinders	-0.78	1.00	0.95	0.84	0.90	-0.50	-0.35
displacement	-0.81	0.95	1.00	0.90	0.93	-0.54	-0.37
horsepower	-0.78	0.84	0.90	1.00	0.86	-0.69	-0.42
weight	-0.83	0.90	0.93	0.86	1.00	-0.42	-0.31
acceleration	0.42	-0.50	-0.54	-0.69	-0.42	1.00	0.29
year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1.00

Perceba que a correlação entre as variáveis explicativas e a variável resposta varia entre moderada e forte, percebe-se também uma correlação forte entre algumas covariáveis, o que pode vir a gerar problemas de multicolinariedade no futuro, essas afirmações serão testadas na modelagem dos dados.

Por fim, retornando ao fato que a variável *origin* é do tipo fator, podemos identificar a distribuição dos dados agrupados pela mesma visualizada pelos bloxplots como segue.

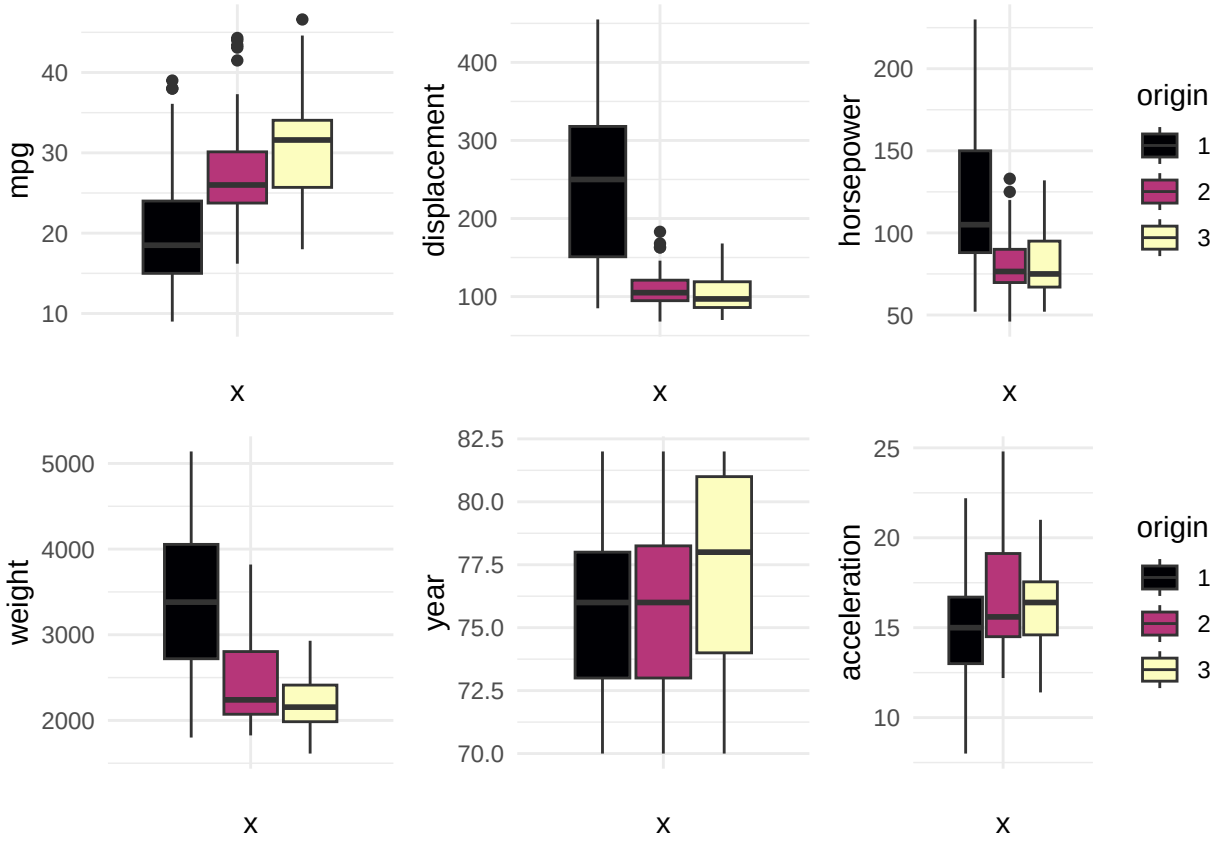


Figure 2: BoxPlots separados por origem do carro

Perceba que a variável parece influenciar sim as outras, ou seja, diferentes resultados dependendo da origem do carro. Note pore xemplo as variáveis *displacement* e *horsepower*, para carros de origem americana temos uma grande dispersão dos dados enquanto que para as demais origens, dados com uma menor variabilidade, esse fator também segue para a variável *weight*. Para variáveis como *mpg* e *accelaration* notamos uma diferença de valores médios para cada um dos grupos, porém uma variábilidade não tão discrepante como para as outras variáveis.

Modelagem

A etapa de modelagem da análise de regressão tem como objetivo construir um modelo estatístico que explique a relação entre as variáveis dependentes e independentes. No caso dos dados Auto, o objetivo é construir um modelo que explique a relação entre a variável MPG (milhas por galão) e as variáveis independentes (ou preditoras) que possam influenciar seu valor.

Considerando as variáveis descritas, de forma incial realizaremos uma seleção das variáveis que irão no modelo, utilizando o método backward e forward stepwise para seleção de variáveis. De forma incial será treinado um modelo formado pelas variáveis e suas possíveis interações, usando a partir deste modelo a técnica backward, usando como medida o AIC e p-valor. Obttemos as variáveis preditoras: Cylinders, displacement, horsepower, weight, accelaration, year, origin e as interações cylinder-acceleration, displacement-weight, displacement-origin, horsepower-year, weight-origin, acceleration-year e acceleration-origin, apresentando um R-ajustado de 0,8873. Note porém que ainda sim o modelo não parece seguir o princípio da parcimonia.

Usando o método forward pelas métricas aplicadas ao backward selection, obtemos as mesmas variáveis independentes do método anterior. A partir desse modelo, usando o nível de significância das variáveis, reduzimos ao modelo:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9+\beta_{10}X_{10}+\beta_{11}X_{11}+\beta_{12}X_{12}+\beta_{13}X_{13}+\epsilon$$

Onde temos respectivamente:

- Y Mpg;
- X_1 cylinders;
- X_1 displacement;
- X_1 horsepower;
- X_1 weight;
- X_1 acceleration;
- X_1 origin (1) ;
- X_1 origin (2);
- X_1 cylinders:acceleration;
- X_1 displacement:weight;
- X_1 horsepower:year;
- X_1 acceleration:year;
- X_1 acceleration:origin(1);
- X_1 acceleration:origin(2).

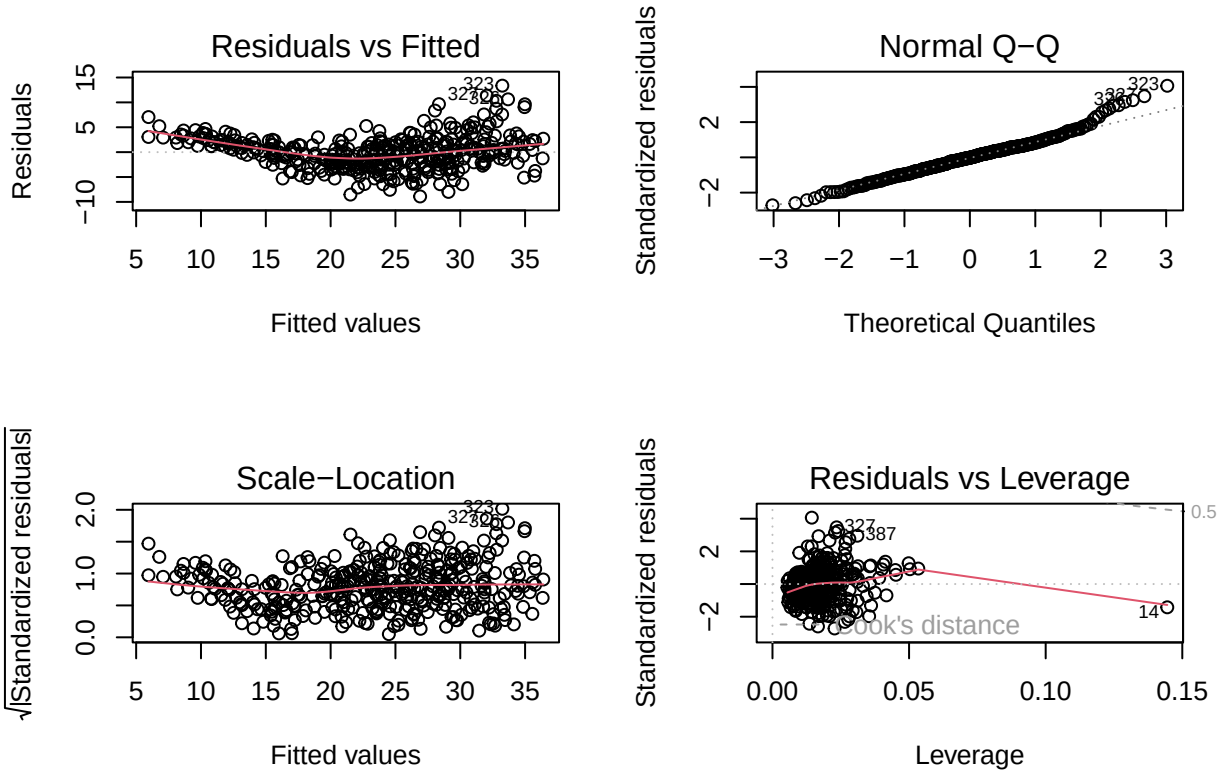
A partir disso, iremos comparar esse modelo com um modelo mais simples, sem interações e aplicado ao método forward selection, com isso obtemos as variáveis independetes: weight, year, origin, displacement e horsepower. Usando método anova para comparação dos modelos, obtemos uma estatistica F de 31.023 , a 7 graus de liberdade, rejeitamos a hipótese de acrescimo das variáveis com interações. Ficamos com o modelo final de:

$$mpg = \beta_0 + \beta_1 weight + \beta_2 year + \beta_3 origin_1 + \beta_4 origin_2 + \beta_5 displacement + \beta_6 horsepower + \epsilon$$

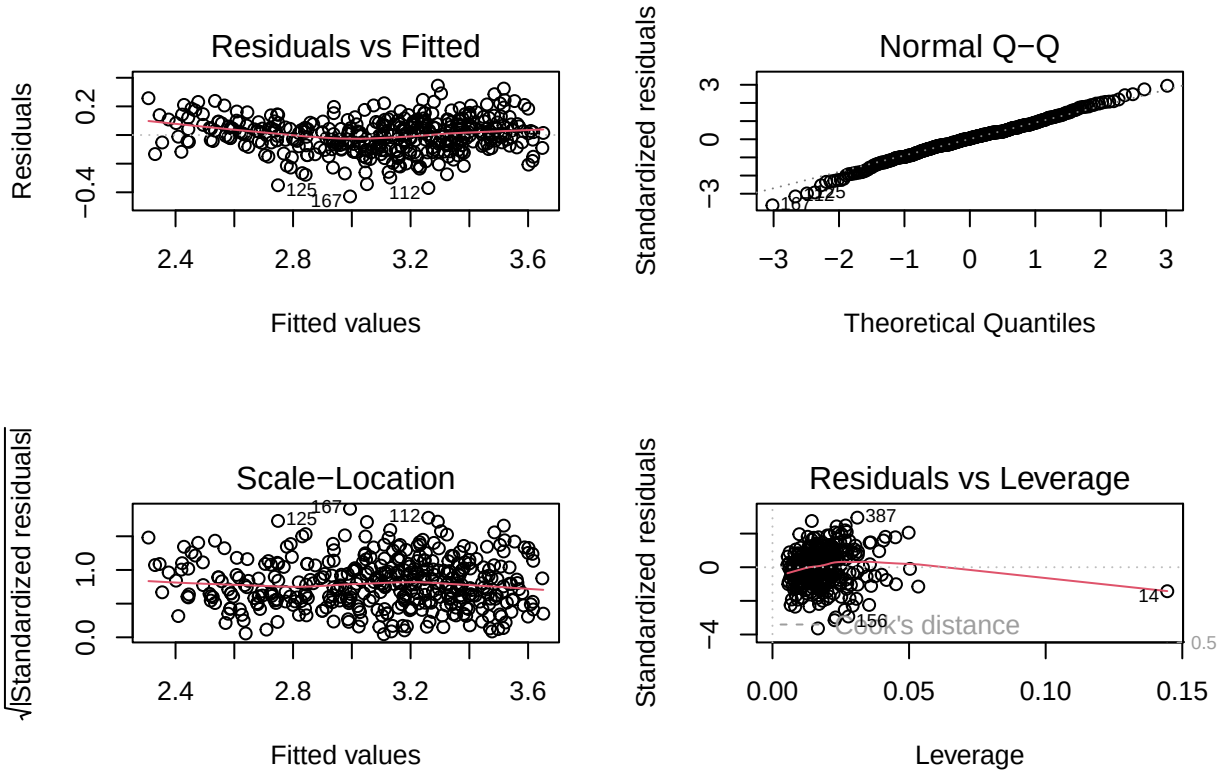
Com um R-Ajustado de 0.8796, um valor ligeiramente inferior ao modelo anterior porém com número consideravelmente inferior de covariáveis, obedecendo o principio da parcimonia então, ficaremos com o último modelo. Observe abaixo o sumário do modelo em questão.

```
##
## Call:
## lm(formula = (mpg) ~ weight + year + origin + displacement +
##     horsepower, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9034 -2.1241 -0.0596  1.8949 13.3558
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.571e+01  4.114e+00  -3.819 0.000156 ***
## weight       -6.565e-03  5.734e-04 -11.449  < 2e-16 ***
## year          7.749e-01  5.171e-02  14.986  < 2e-16 ***
## origin(1)    -1.789e+00  3.212e-01  -5.571 4.77e-08 ***
## origin(2)     8.065e-01  3.299e-01   2.444 0.014961 *
## displacement  1.555e-02  5.760e-03   2.699 0.007253 **
## horsepower   -2.304e-02  1.072e-02  -2.149 0.032266 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## s: 3.311 on 385 degrees of freedom
## Multiple R-squared: 0.8228,
## Adjusted R-squared:  0.82
## F-statistic: 297.9 on 6 and 385 DF,  p-value: < 2.2e-16
```

Se verificarmos os plots do modelo, vemos que a hipótese de normalidade dos resíduos não esta completamente sendo seguida.



Vemos então, que ao aplicar o logaritmo na variável resposta, vemos uma considerável melhora no ajuste do modelo, segue então o diagnóstico do modelo final:



Obtendo então o sumario como:

```
##
## Call:
## lm(formula = log(mpg) ~ weight + year + origin + displacement +
##     horsepower, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42900 -0.07014  0.00600  0.07341  0.34474
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.672e+00  1.477e-01  11.317  < 2e-16 ***
## weight       -2.741e-04  2.059e-05 -13.314  < 2e-16 ***
## year          3.066e-02  1.856e-03  16.514  < 2e-16 ***
## origin(1)    -5.327e-02  1.153e-02  -4.619 5.26e-06 ***
## origin(2)     2.983e-02  1.185e-02   2.519  0.01219 *
## displacement  3.690e-04  2.068e-04   1.785  0.07513 .
## horsepower   -1.292e-03  3.850e-04  -3.356  0.00087 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## s: 0.1189 on 385 degrees of freedom
## Multiple R-squared: 0.8796,
## Adjusted R-squared: 0.8778
## F-statistic:   469 on 6 and 385 DF,  p-value: < 2.2e-16
```

Verificando assim então, a relação positiva entre year, displacement e origen 1 para a variável resposta e uma relação negativa com as demais.