Análise e Modelagem de Preços de Carros Usados

Gustavo Almeida Silva

Universidade Federal de Juiz de Fora

07/07/2023

O Mercado de Carros Usados

- Acessibilidade
- Sustentabilidade

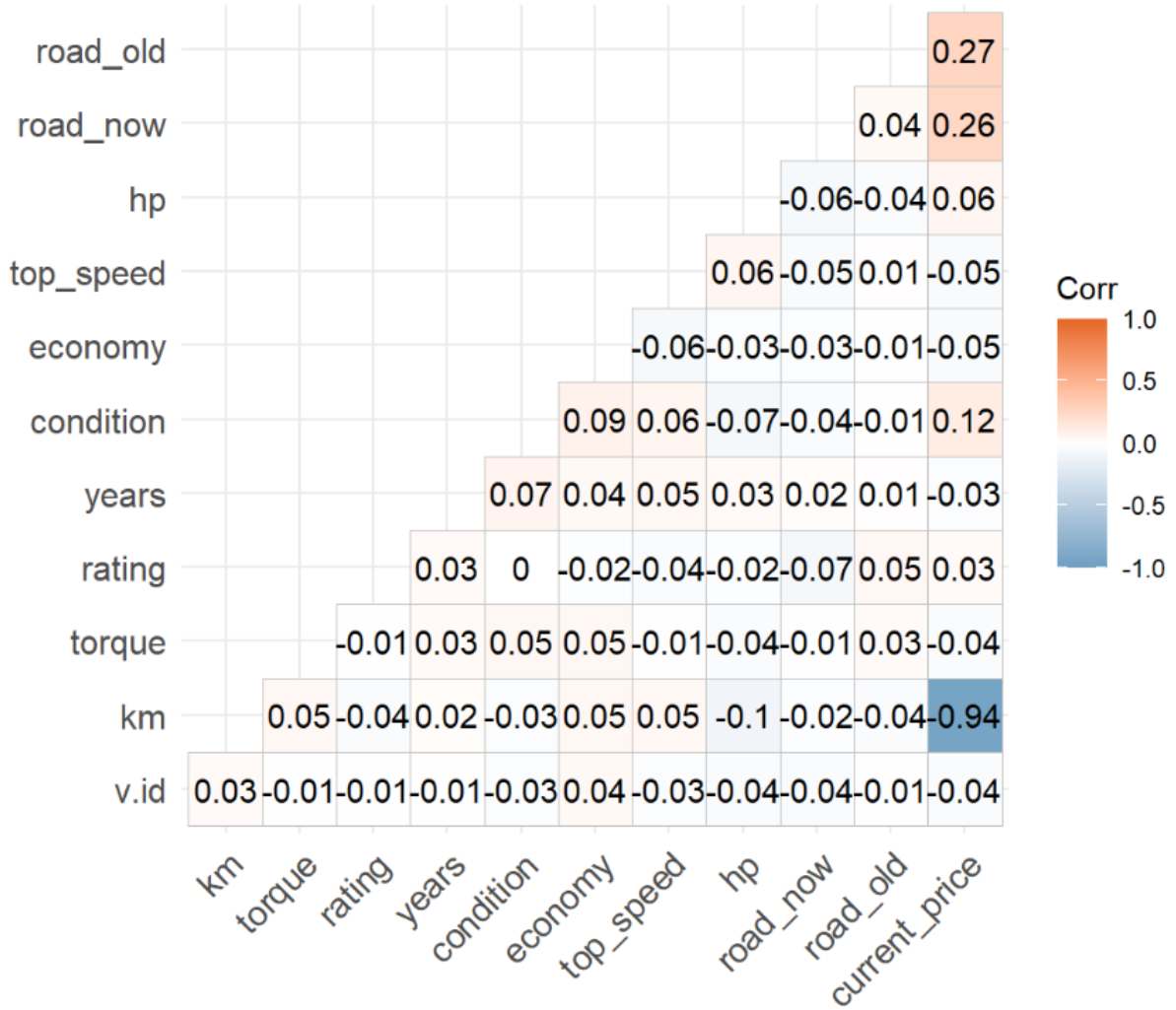
Dados Utilizados

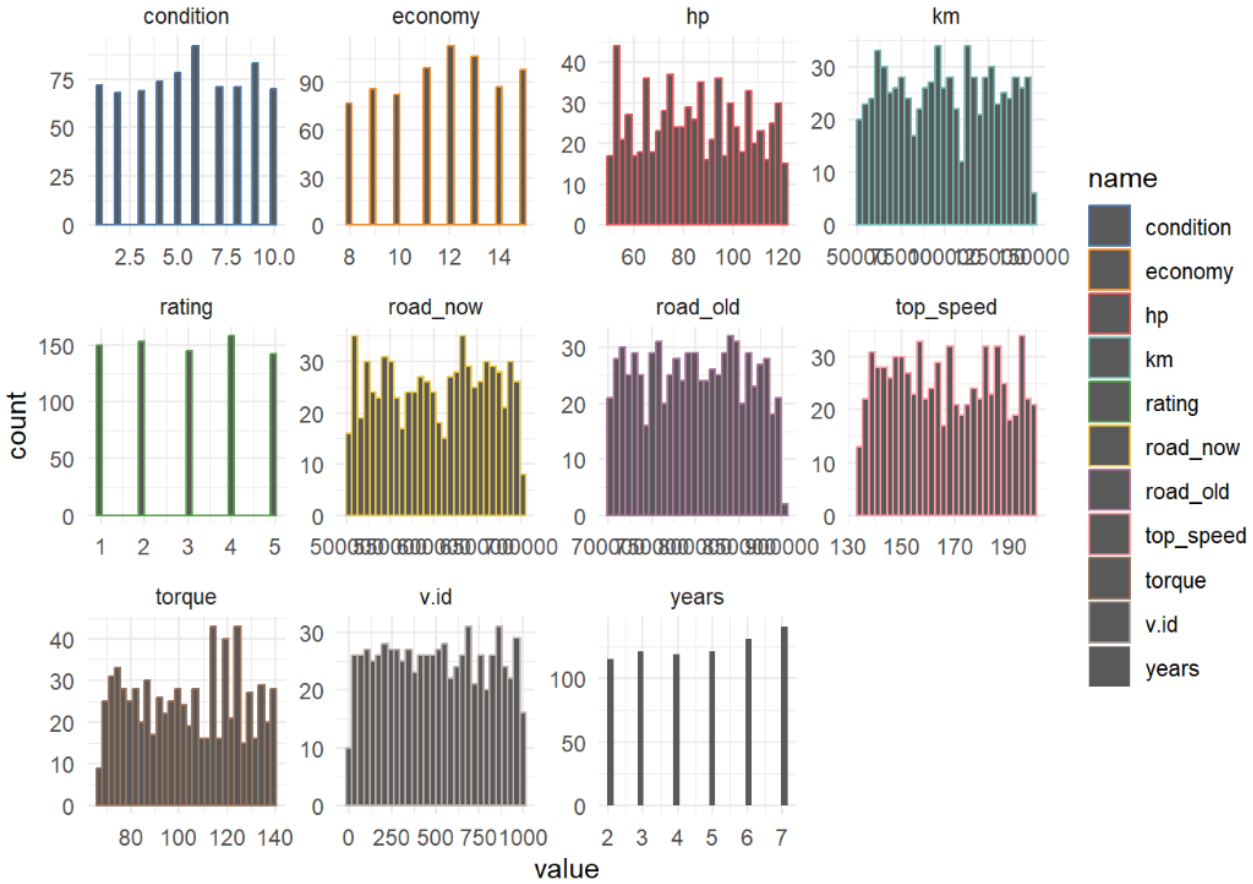
12 variáveis e 1000 observações

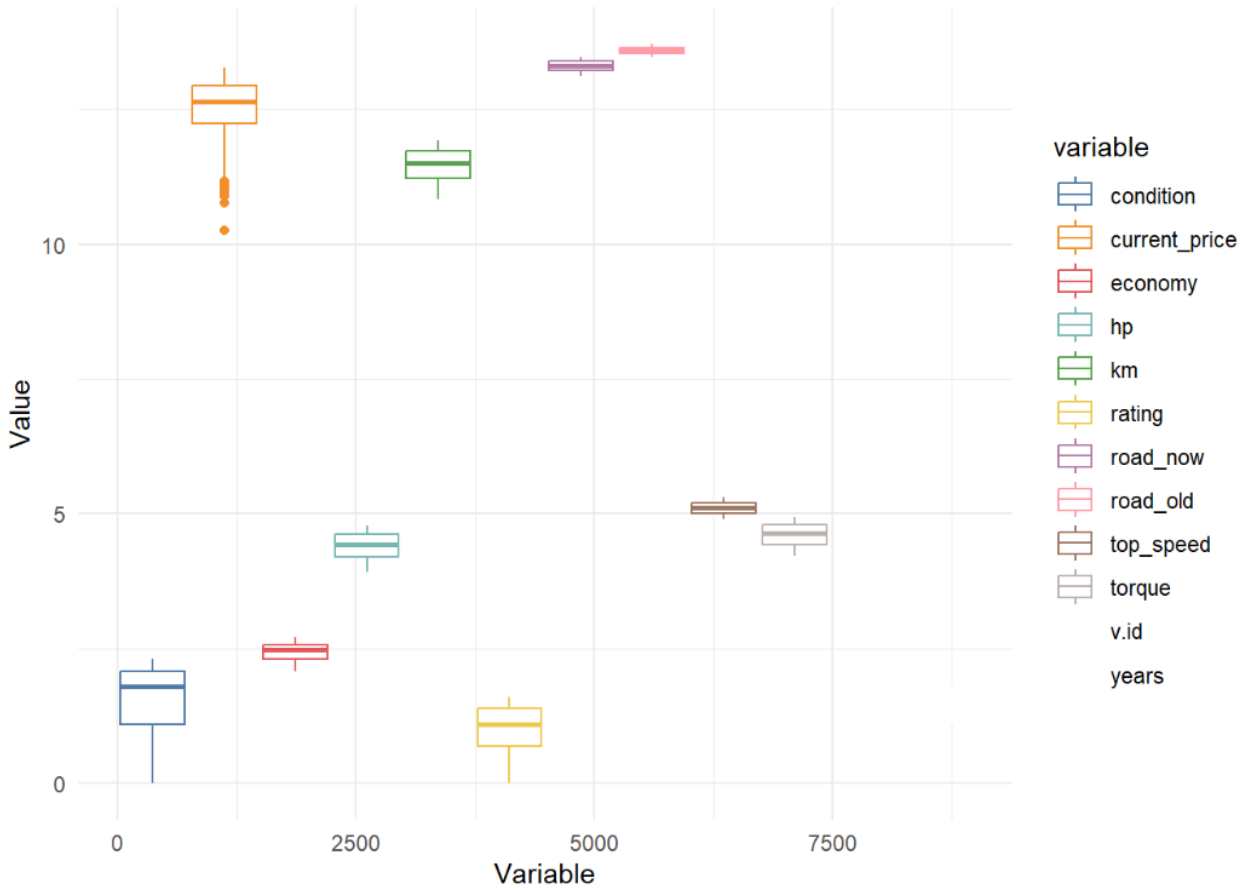
- **v.id**
- **road_old**
- **road_now**
- **years**
- **km**
- **rating**
- **condition**
- **economy**
- **top_speed**
- **hp**
- **torque**
- **current_price**

Separação e Análise Exploratória dos Dados

- Divisão Treino e Teste (0.75, 0.25)
- Ausência de Valores Faltantes




condition
economy
hp
km
rating
road_now
road_old
top_speed
torque
v.id
years
count
value
name
condition
economy
hp
km
rating
road_now
road_old
top_speed
torque
v.id
years


Value
10
5
0
0
2500
5000
7500
Variable
variable
condition
current_price
economy
hp
km
rating
road_now
road_old
top_speed
torque
v.id
years

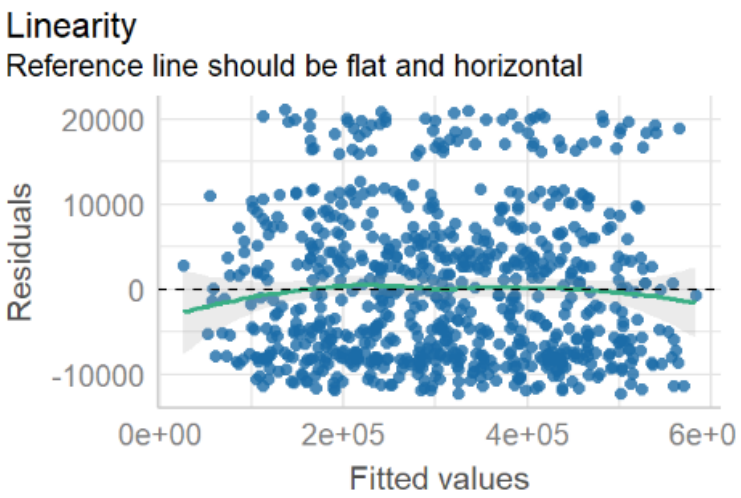
Modelos

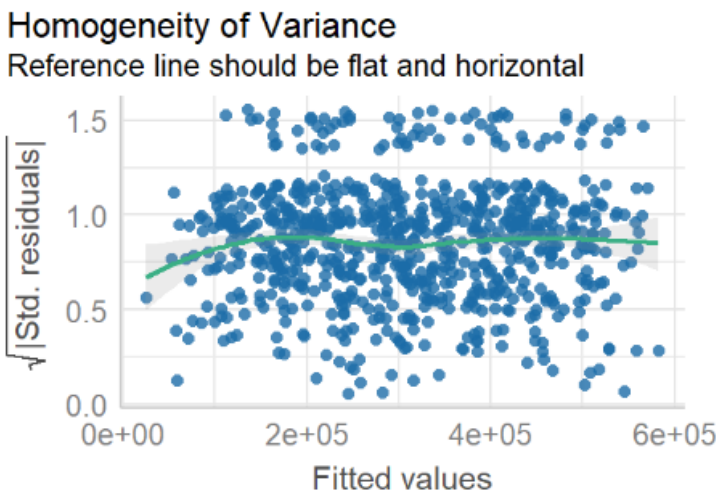
- Saturado
- Correlacional

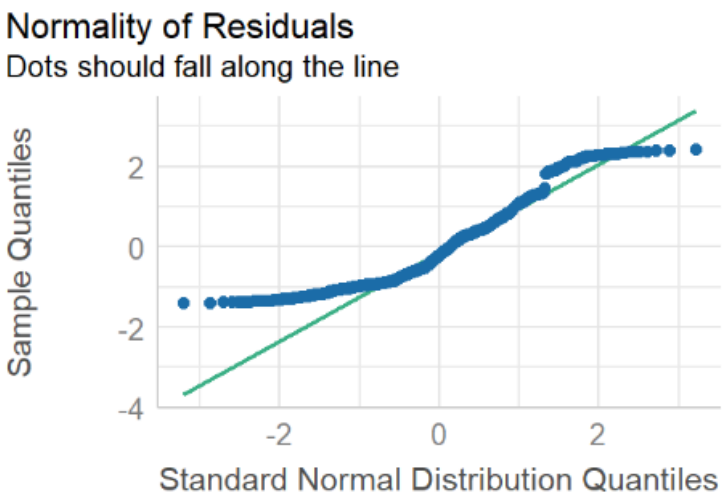
	Saturado	Correlacional
(Intercept)	-14651.9*	20605.4*
	p = <0.1	p = <0.1
road_now	0.5***	0.5***
	p = <0.1	p = <0.1
road_old	0.5***	0.5***
	p = <0.1	p = <0.1
years	-1595.4***	-1063.3**
	p = <0.1	p = <0.1
km	-4.0***	-4.0***
	p = <0.1	p = <0.1
rating	285.8	
	p = 0.2	
condition	4629.4***	
	p = <0.1	
economy	41.7	
	p = 0.8	
top_speed	-7.9	

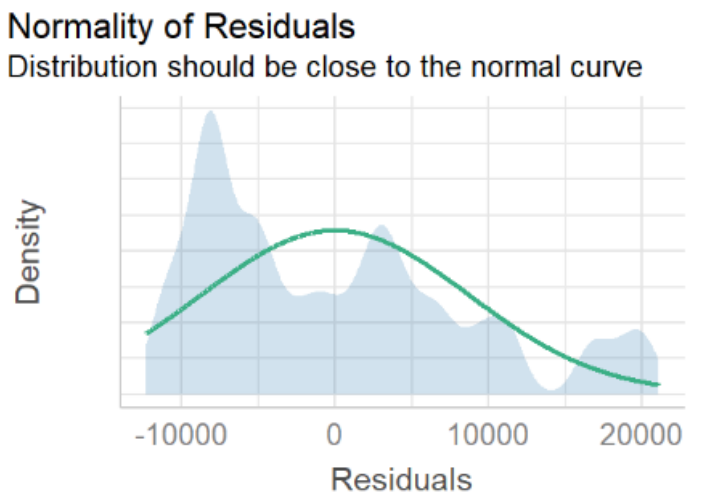
Num.Obs.	748	748
R2	0.995	0.985
R2 Adj.	0.995	0.984
AIC	15720.9	16585.0
BIC	15776.3	16612.7
Log.Lik.	-7848.427	-8286.511
RMSE	8722.06	15666.58

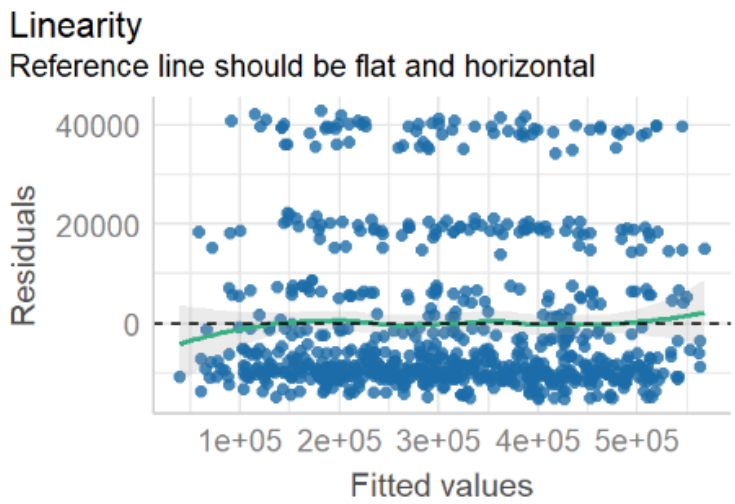
Residuos

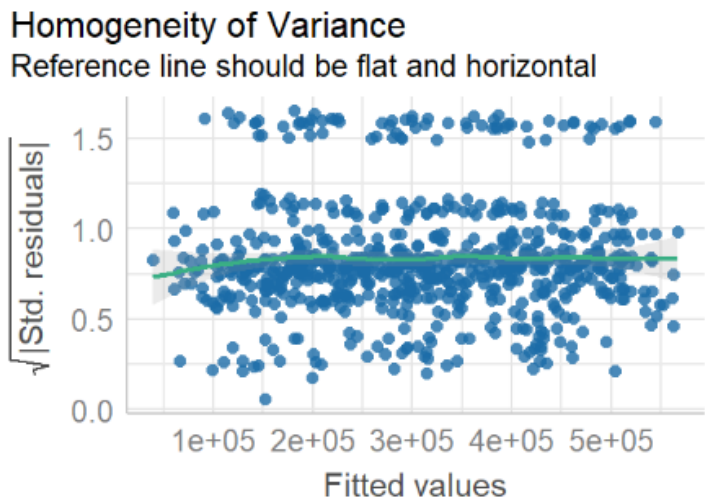

Linearity
Reference line should be flat and horizontal
Residuals
20000
10000
0
-10000
0e+00
2e+05
4e+05
6e+0
Fitted values

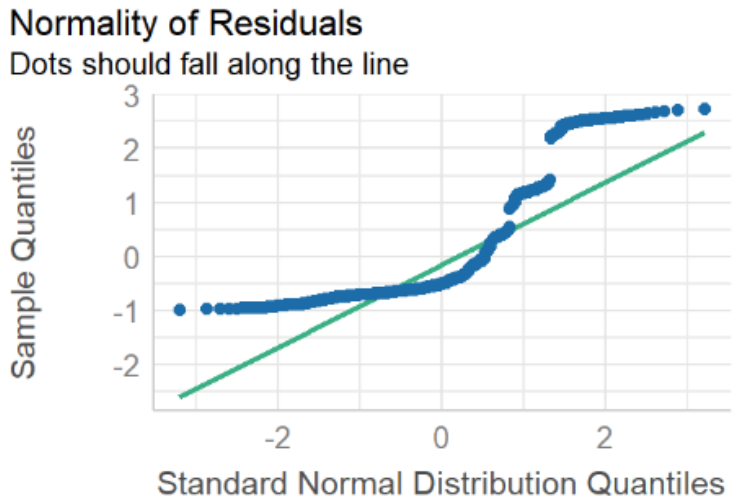

Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
1.5
1.0
0.5
0.0
0e+00
2e+05
4e+05
6e+05
Fitted values

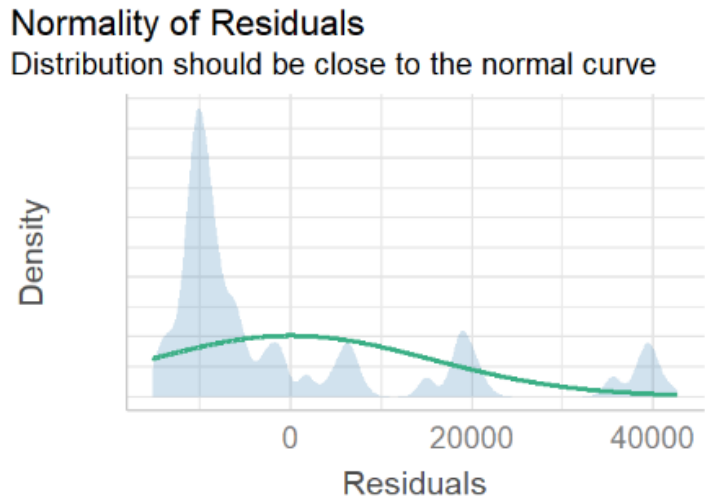

Normality of Residuals
Dots should fall along the line
Sample Quantiles
2
0
-2
-4
-2
0
2
Standard Normal Distribution Quantiles


Normality of Residuals
Distribution should be close to the normal curve
Density
-10000
0
10000
20000
Residuals


Linearity
Reference line should be flat and horizontal
Residuals
40000
20000
0
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values


Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
1.5
1.0
0.5
0.0
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values


Normality of Residuals
Dots should fall along the line
Sample Quantiles
3
2
1
0
-1
-2
-2
0
2
Standard Normal Distribution Quantiles

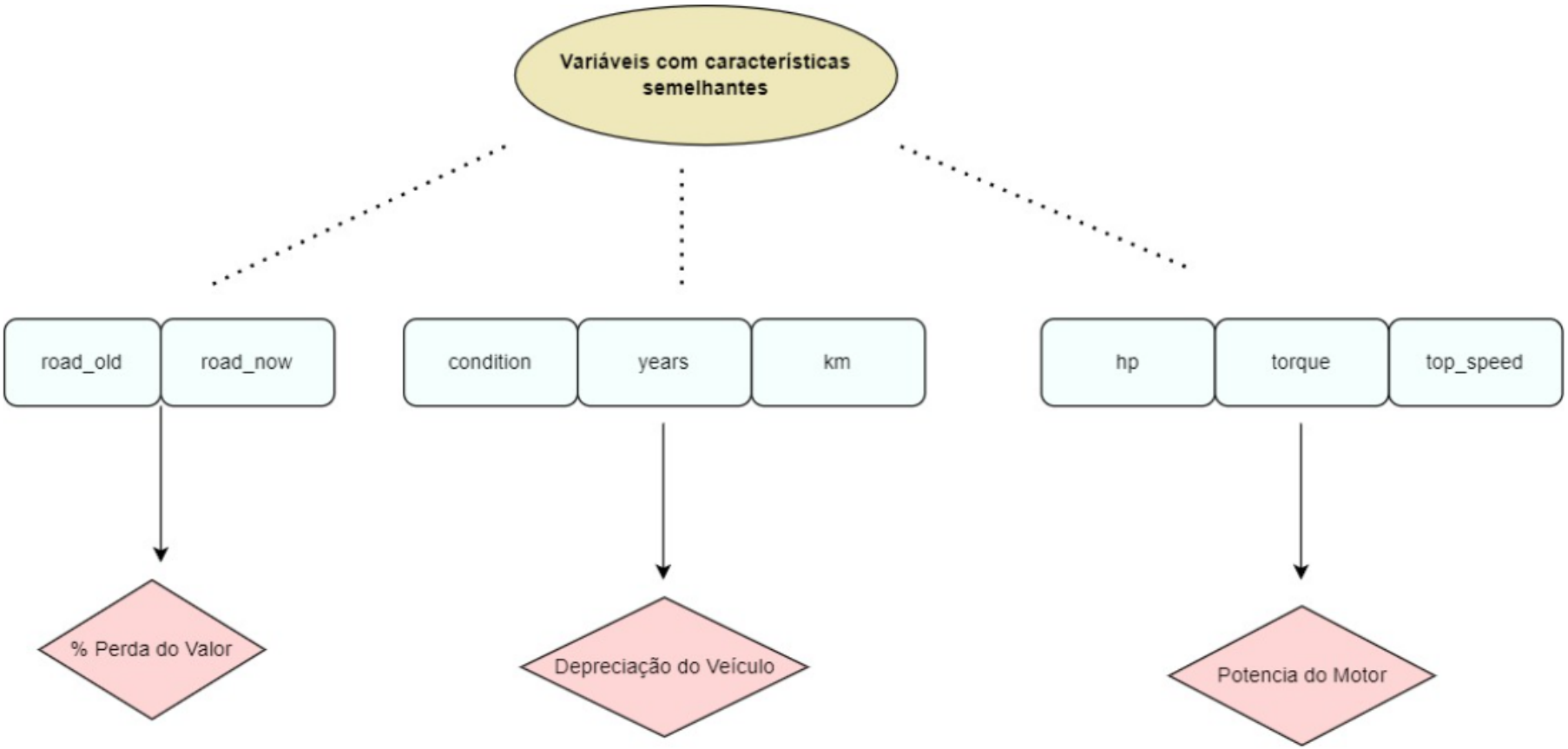

Normality of Residuals
Distribution should be close to the normal curve
Density
0
20000
40000
Residuals

Teste	Modelo.Sat.P.valor	Modelo.Corr.P.valor
Shapiro	0	0
Cramer	0	0
Lilliefors	0	0

- Rejeição da Hipótese de Normalidade
- Heterogeneidade da Variância

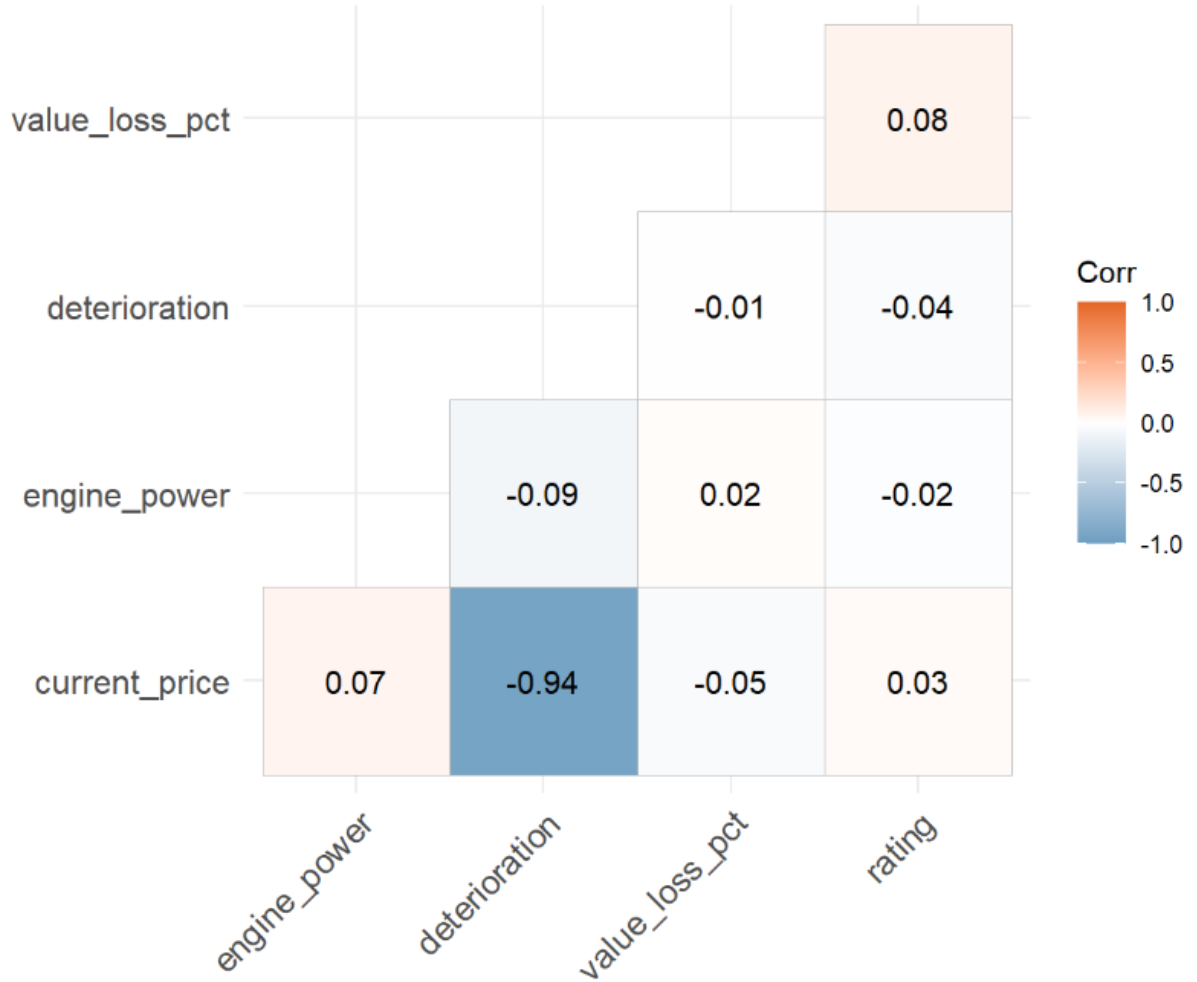
Engenharia de Características

- Criação de novas variáveis via transformação
- Tentativa de melhorar o modelo


Variáveis com características semelhantes
road_old
road_now
condition
years
km
hp
torque
top_speed
% Perda do Valor
Depreciação do Veículo
Potencia do Motor

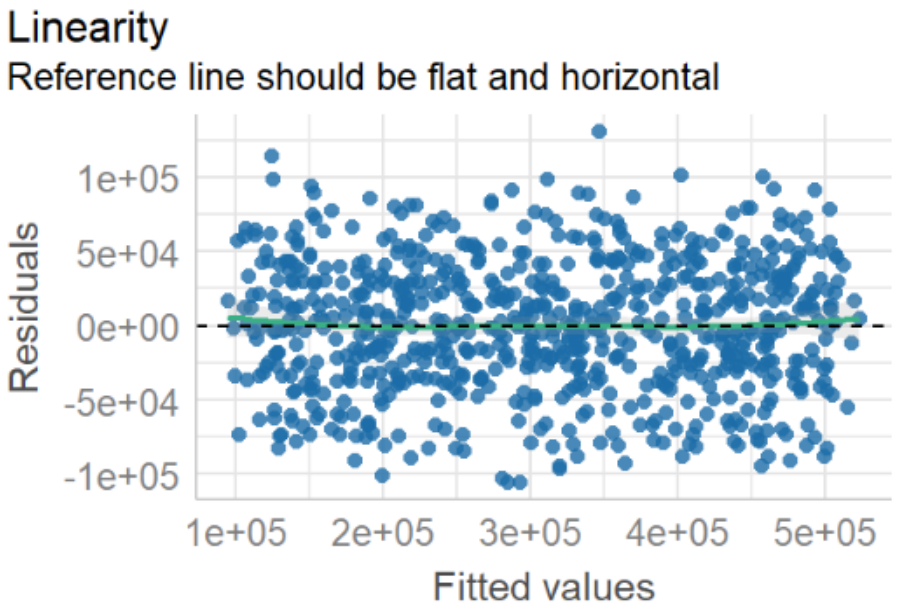
Transformações utilizadas

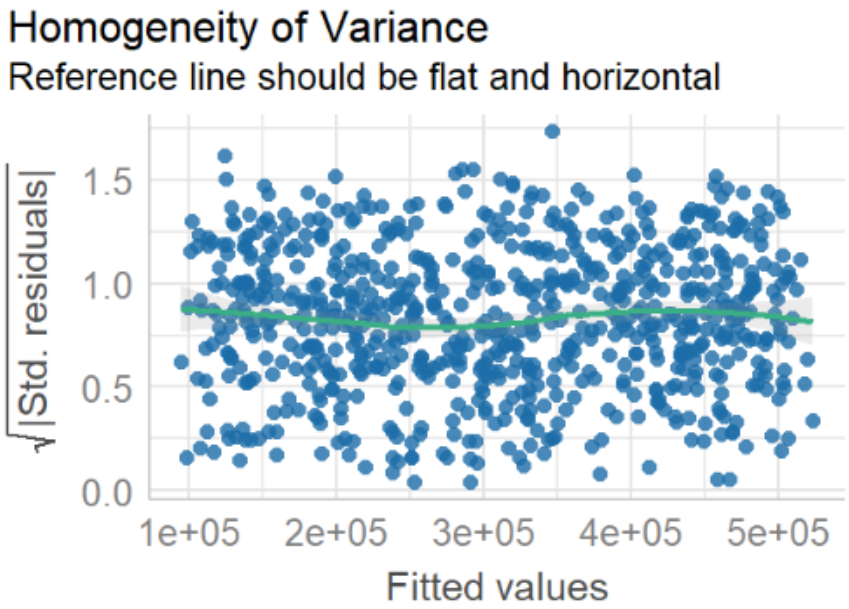
- **% Perda de Valor** $= \frac{(ValorInicial - ValorAtual) \times 100}{ValorInicial}$
- **Deterioração do Veículo** $= Media(condition,\ years,\ km)$
- **Potencia do Motor** $= Media(top\ speed^2,\ hp^5,\ torque)$

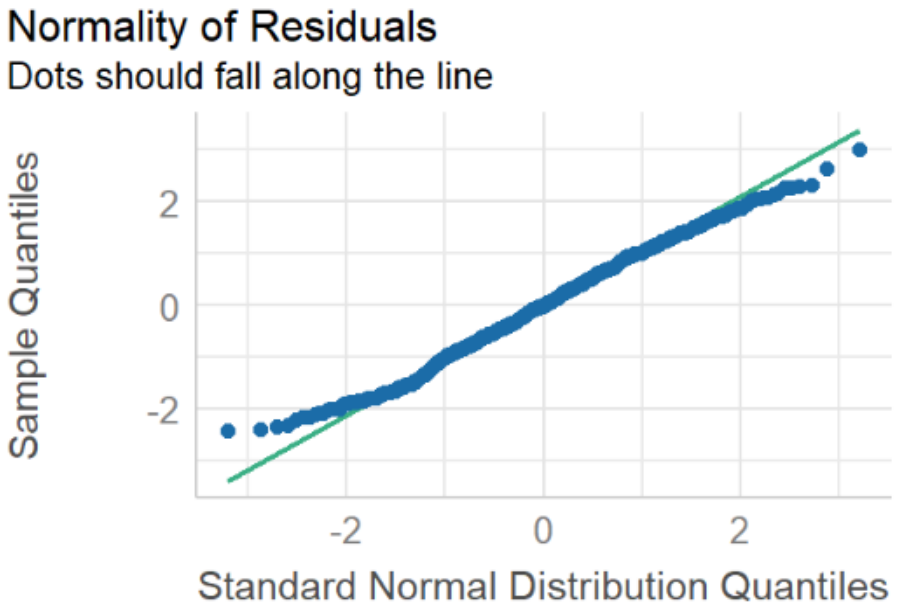


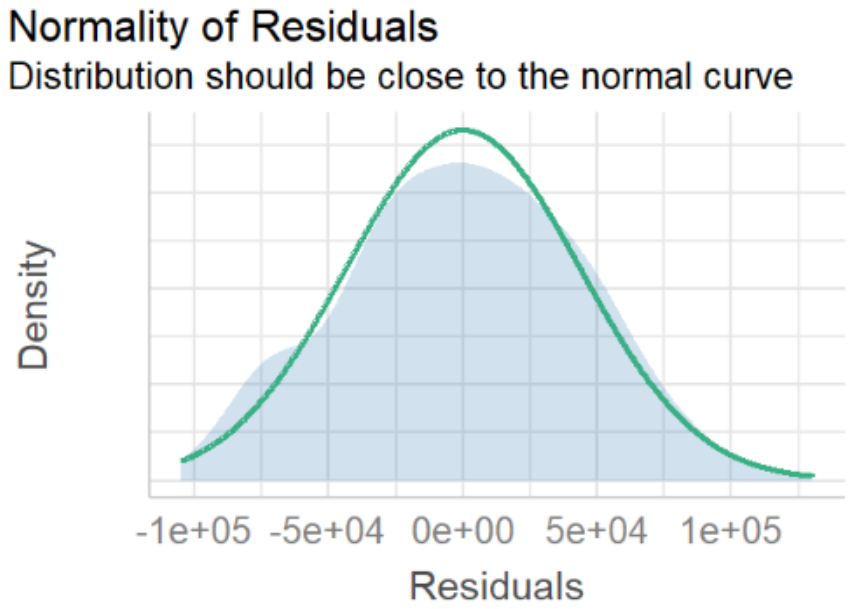
- Modelo: $current\_price = \beta_0 + \beta_1\ value\_loss\_pct + \beta_2\ deterioration + \beta_3\ engine\_power + E$

	Features
(Intercept)	739155.0***
	p = <0.1
value_loss_pct	-771.7***
	p = <0.1
deterioration	-12.3***
	p = <0.1
engine_power	0.0
	p = 0.1
Num.Obs.	748
R2	0.880
R2 Adj.	0.879
AIC	18115.6
BIC	18138.7
Log.Lik.	-9052.782
RMSE	43639.20

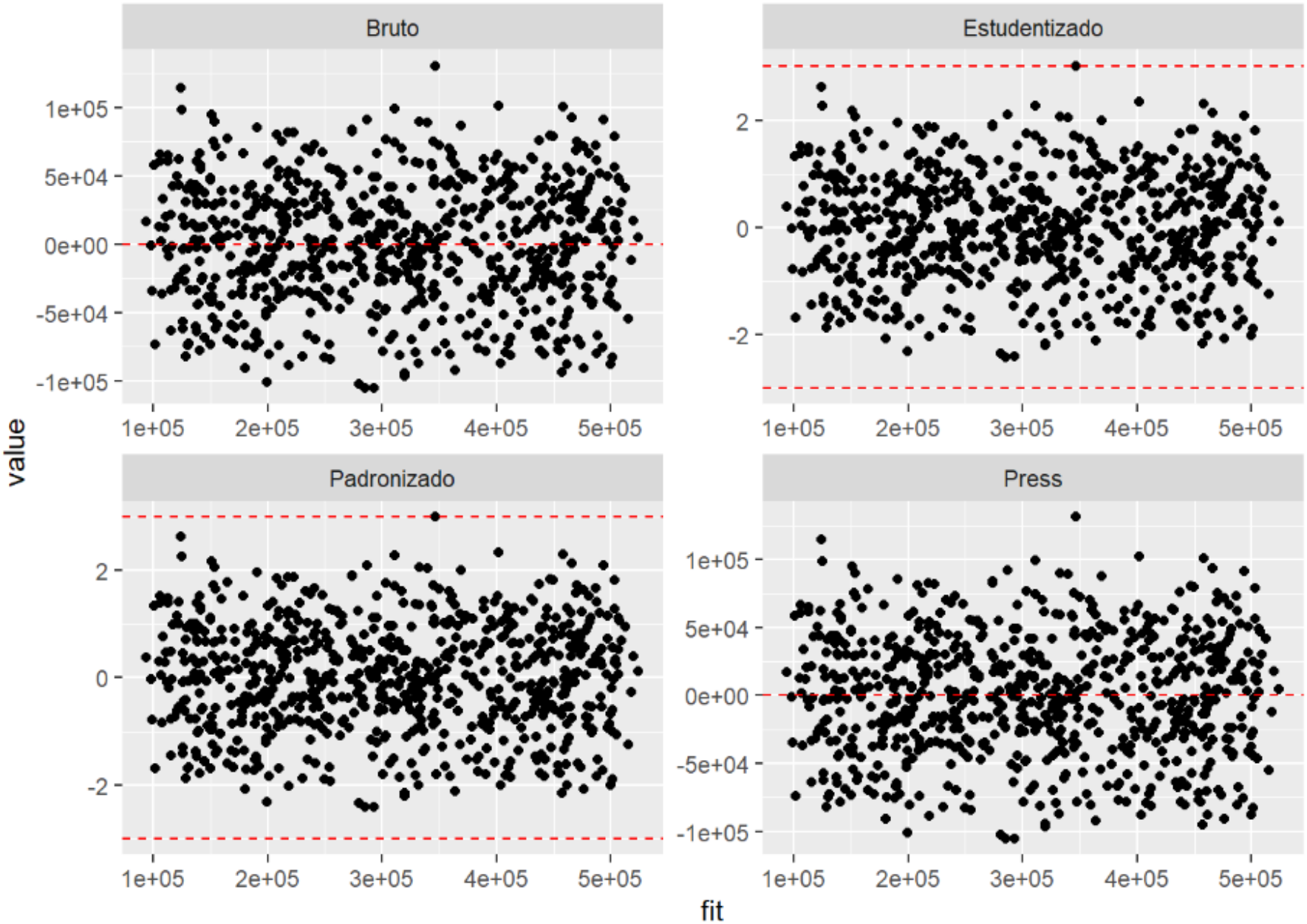

Linearity
Reference line should be flat and horizontal
Residuals
1e+05
5e+04
0e+00
-5e+04
-1e+05
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values


Homogeneity of Variance
Reference line should be flat and horizontal
√|Std. residuals|
1.5
1.0
0.5
0.0
1e+05
2e+05
3e+05
4e+05
5e+05
Fitted values

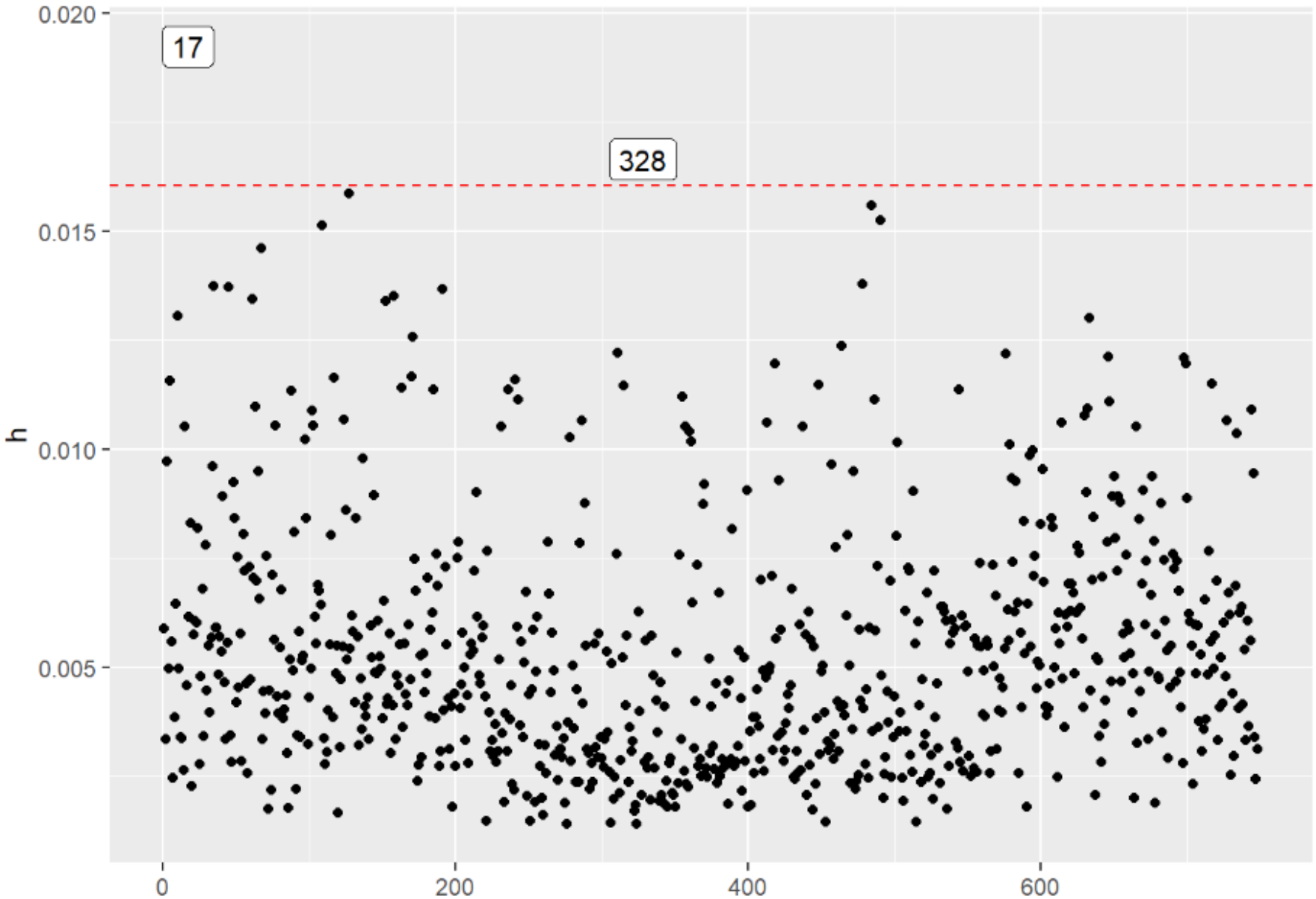

Normality of Residuals
Dots should fall along the line
Sample Quantiles
2
0
-2
-2
0
2
Standard Normal Distribution Quantiles

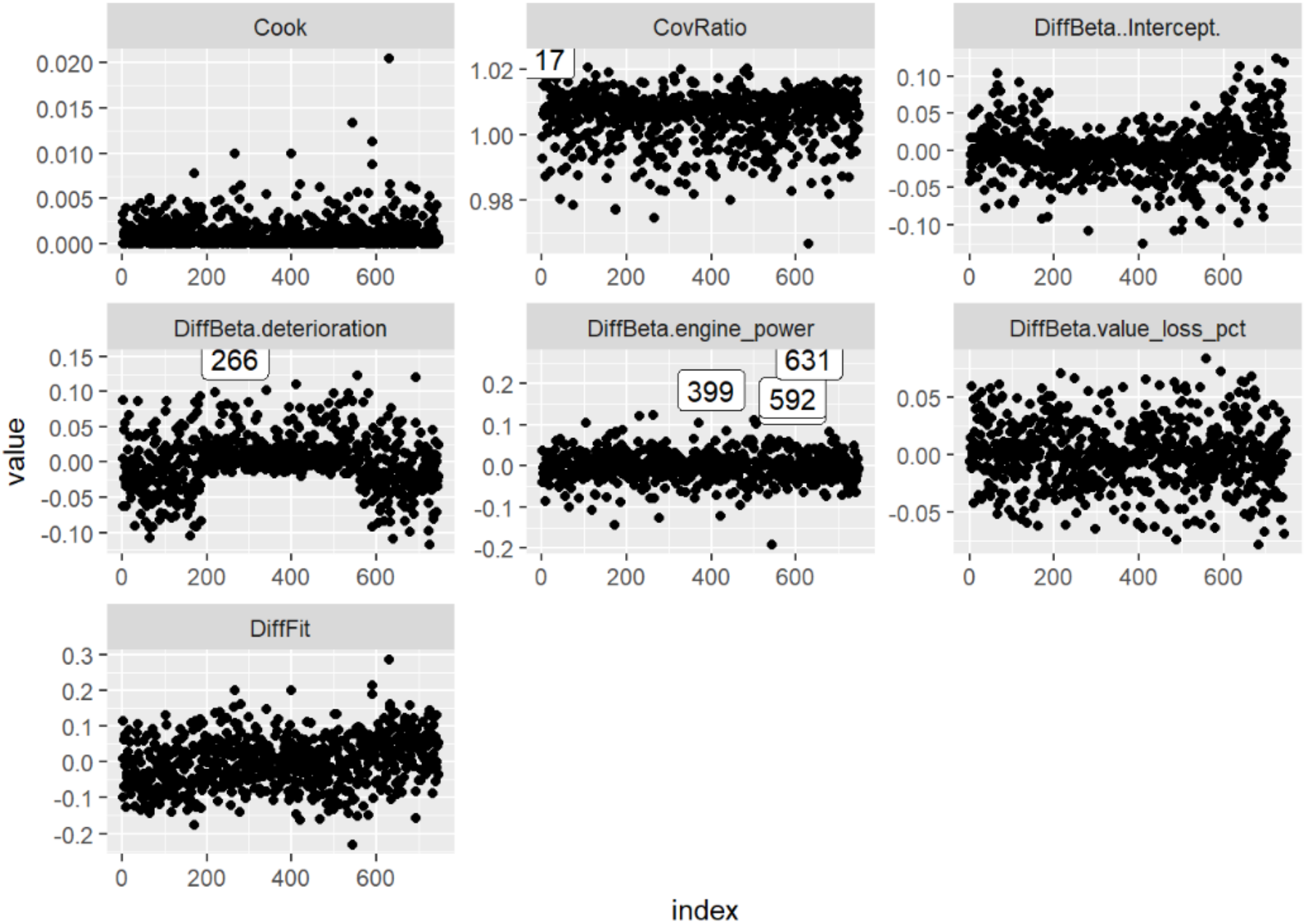

Normality of Residuals
Distribution should be close to the normal curve
Density
-1e+05
-5e+04
0e+00
5e+04
1e+05
Residuals

Teste	Modelo.Feat.P.valor
Shapiro	0.2048031
Cramer	0.3276374
Lilliefors	0.4188786


Bruto
Estudentizado
Padronizado
Press
value
fit
1e+05
5e+04
0e+00
-5e+04
-1e+05
2
0
-2
1e+05
2e+05
3e+05
4e+05
5e+05

Análise de Diagnósticos



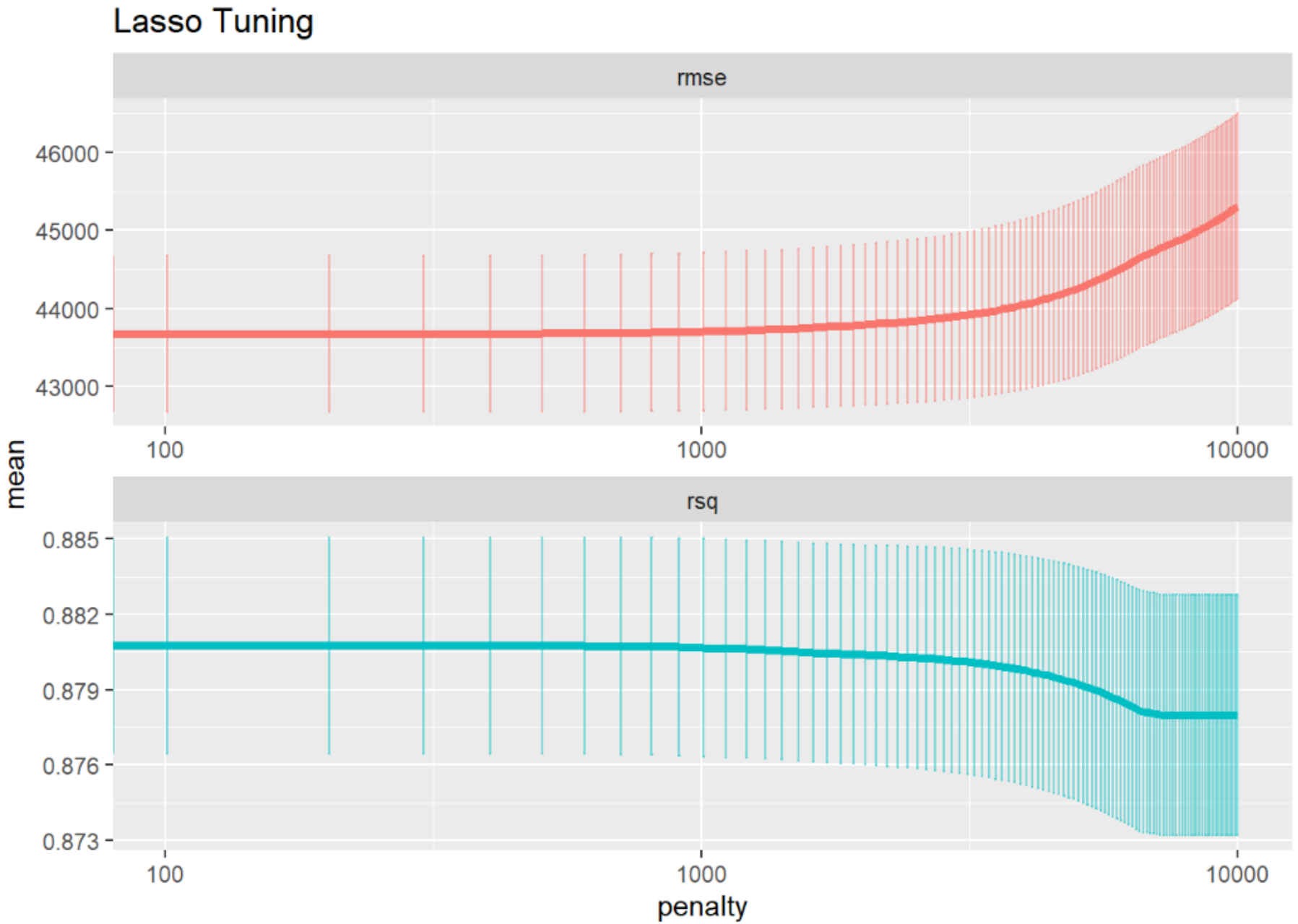

Cook
CovRatio
DiffBeta..Intercept.
DiffBeta.deterioration
DiffBeta.engine_power
DiffBeta.value_loss_pct
DiffFit
17
266
399
592
631
value
index

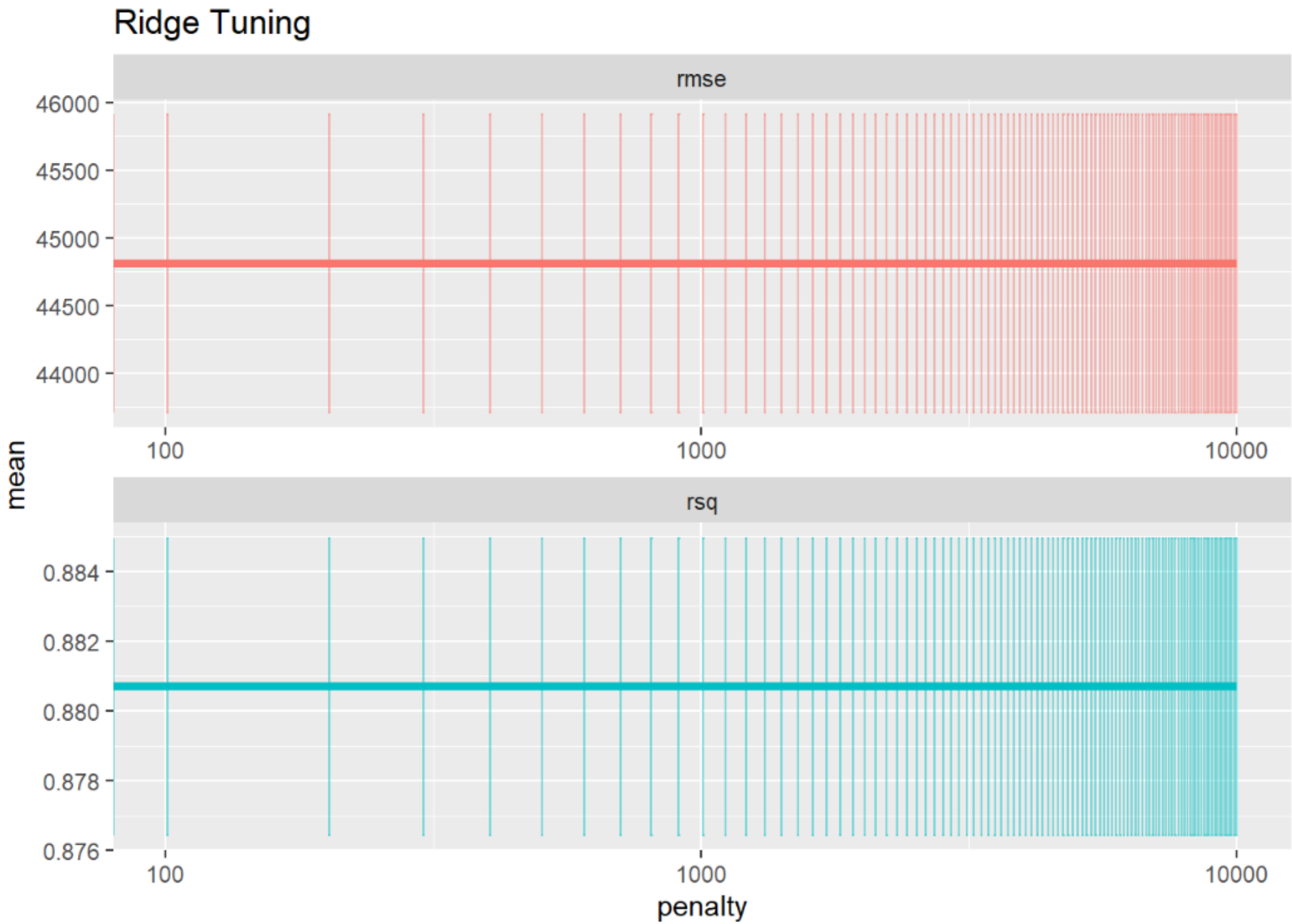
Multicolinearidade

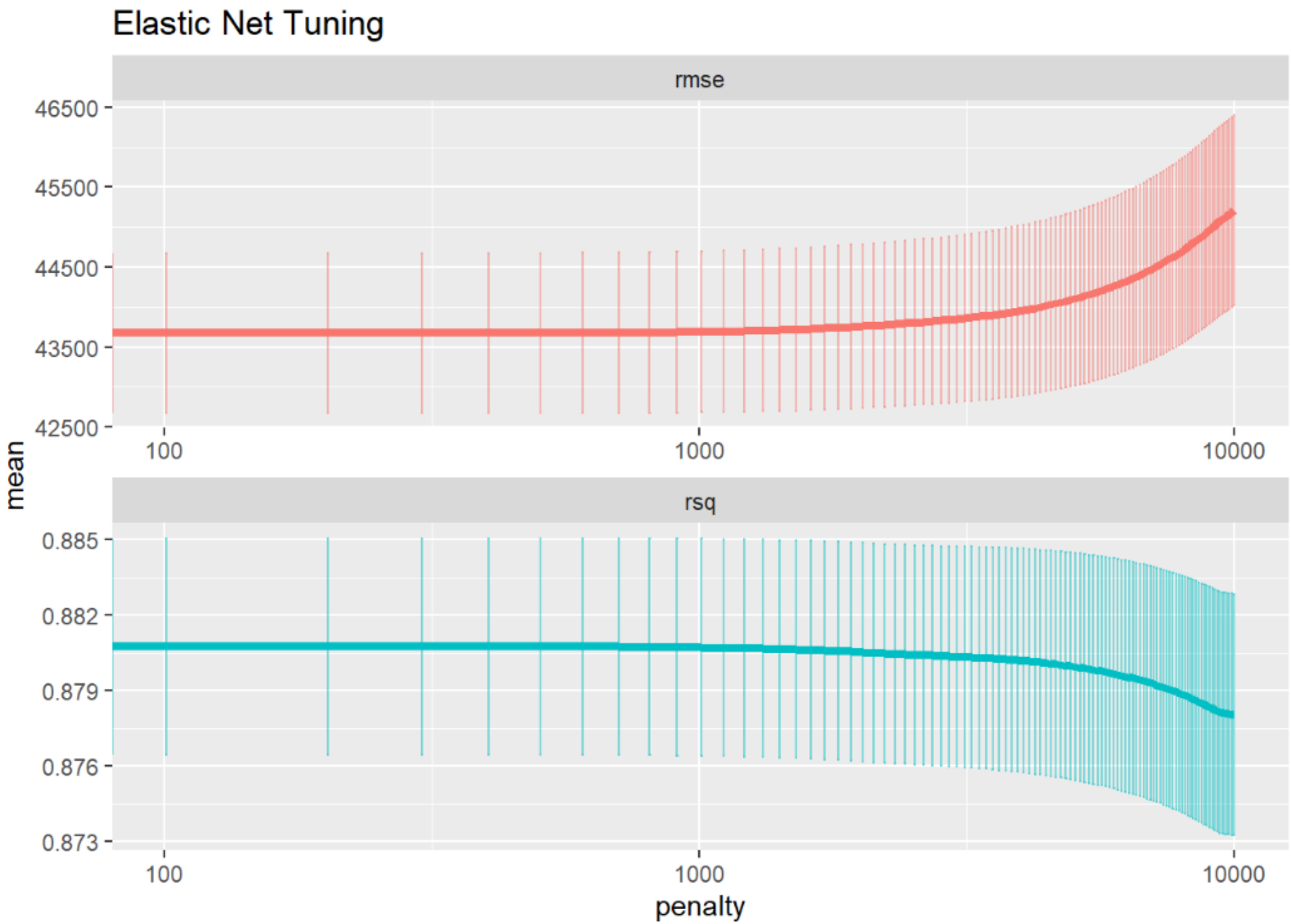
	VIF_Values
value_loss_pct	1.000291
deterioration	1.008633
engine_power	1.008797

Regularização

- Busca penalizar variávies colineares
- Modelo não apresenta multicolinearidade, e portanto não espera-se melhora
- K-fold e Leave One Out para robuste de métricas
- **Lasso (L1)**
- **Ridge (L2)**
- **Elastic Net (Mistura 0.7 de penalty L1 e 0.3 de Penalty L2)**

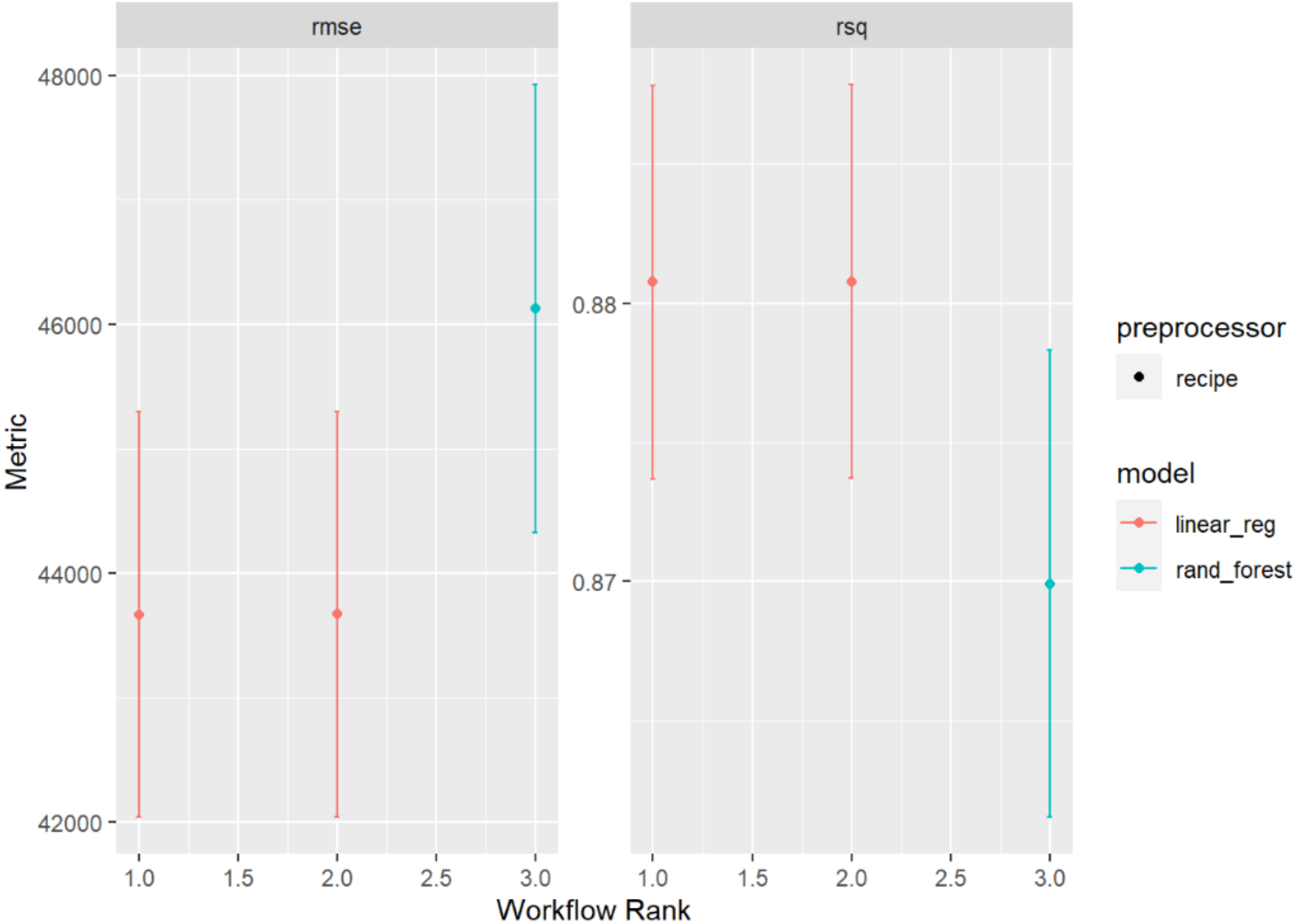

Lasso Tuning
rmse
46000
45000
44000
43000
100
1000
10000
mean
rsq
0.885
0.882
0.879
0.876
0.873
100
1000
10000
penalty


Ridge Tuning
rmse
46000
45500
45000
44500
44000
100
1000
10000
mean
rsq
0.884
0.882
0.880
0.878
0.876
100
1000
10000
penalty


Elastic Net Tuning
rmse
rsq
mean
penalty
46500
45500
44500
43500
42500
0.885
0.882
0.879
0.876
0.873
100
1000
10000

Tópico extra

- Assuntos Comentados em Aula
- Modelo Bayesiano (família Gaussina sem fixação de hiperparâmetros)
- Aprendizado de Máquina (Random Forest sem tuning de hiperparâmetros)
- K-fold e Leave One Out para robuste de métricas


rmse
rsq
48000
46000
44000
42000
0.88
0.87
Metric
1.0
1.5
2.0
2.5
3.0
Workflow Rank
preprocessor
recipe
model
linear_reg
rand_forest

- Métricas iguais e piores do que o modelo linear simples
- Complexidade de Interpretação

Validação

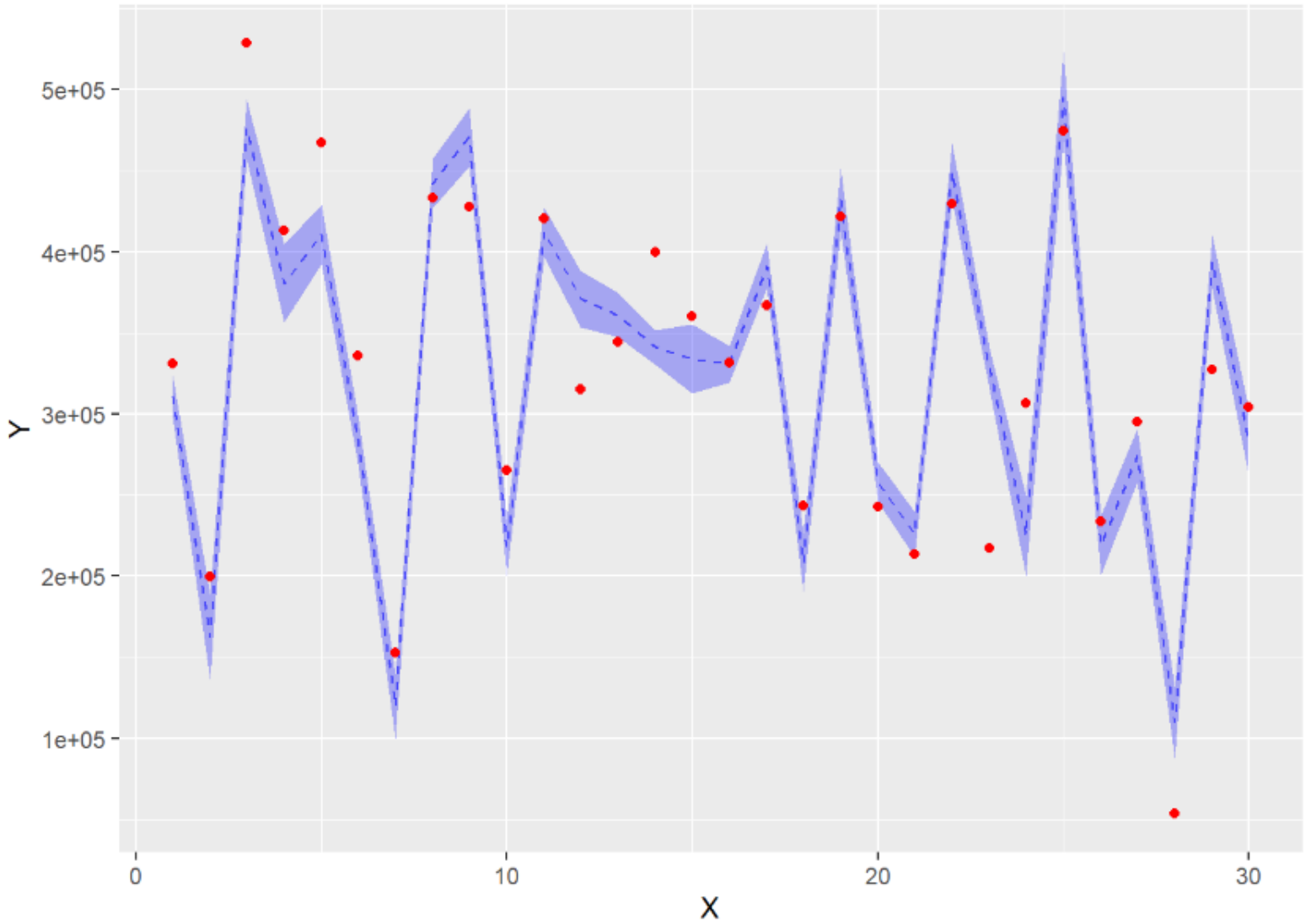
- Melhor Modelo: $current\_price = \beta_0 + \beta_1\ value\_loss\_pct + \beta_2\ deterioration + \beta_3\ engine\_power + E$

	Features
(Intercept)	739155.0***
	p = <0.1
value_loss_pct	-771.7***
	p = <0.1
deterioration	-12.3***
	p = <0.1
engine_power	0.0
	p = 0.1
Num.Obs.	748
R2	0.880
R2 Adj.	0.879
AIC	18115.6
BIC	18138.7
Log.Lik.	-9052.782
RMSE	43639.20

Validação no Conjunto de Teste

Metric	Data_test	Data_Training
rmse	4.438387e+04	4.363920e+04
rsq	8.773829e-01	8.797584e-01

Valores Preditos



Conclusão

- O trabalho utilizou técnicas de regressão linear, como modelos saturados e correlacionais, engenharia de características, análise de diagnósticos e análise de resíduos
- Os resíduos do modelo não rejeitaram a hipótese de normalidade, o que indica um bom ajuste.
- A engenharia de características permitiu criar um modelo com baixa multicolinearidade, conforme indicado pelos baixos valores de VIF
- A aplicação de regularização (L1, L2 e Elastic Net) não trouxe melhorias significativas ao modelo.
- Comparando com outros modelos, como os Bayesianos e Random Forest, o modelo de regressão linear simples obteve resultados melhores e uma boa capacidade de generalização.