VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS

II LABORATORINIS DARBAS

AUTOMOBILIŲ KURO SĄNAUDŲ PROGNOZAVIMAS

MAKSIM ČIŽOV, ANTON CIFIROV

Docentė
Dr. Rūta Levulienė

VILNIUS, 2024

TURINYS

ĮVADAS IR DUOMENYS

Laboratoriniam darbui atlikti pasirinkome viešai prieinamą duomenų rinkinį „Auto MPG“, turinti 398 stebinius ir 8 rodiklius, iš kurių 2 yra kategoriniai (automobilio markė ir kilmės regionas) ir 6 skaitiniai (variklio tūris, variklio cilindrų skaičius, variklio galingumas, automobilio svoris, pagaminimo metai ir greitėjimo lygis), bei atsako kintamąjį „mpg“, nurodanti automobilio kuro sąnaudas. Analizei atlikti nusprendėme pasinaudoti ir palyginti Gama regresijos ir Atvirkštinės Gauso regresijos modelius su skirtingomis jungties funkcijomis tarpusavyje. Modeliui sukurti naudojome R, Python ir SAS programavimo kalbas.

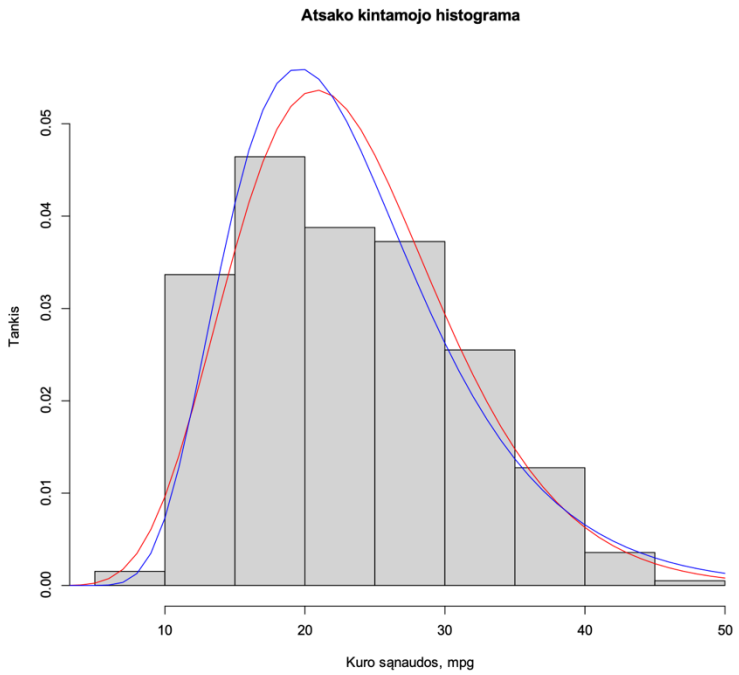
Tikslas - sukruti Gama regresijos arba Atvirkštinės Gauso regresijos modelį, kuris iš turimu požymių prognozuotu automobilio kuro sąnaudas.

Uždaviniai:

1. Atlikti pradinę duomenų analizę
2. Patikrinti kintamųjų multikolinearumą
3. Atlikti išimčių analizę
4. Atrinkti tinkamus regresorius
5. Sudaryti Gama regresijos ir Atvirkštinės Gauso regresijos modelius, jas palyginti ir įvertinti rezultatus
6. Padaryti išvadas

SPRENDIMAS SU R

Nuskaičius duomenis, atlikome pirminę duomenų analizę: panaikinome nulines stebinių reikšmes, bei, pasinaudojus Kolmogorovo-Smirnovo testu, patikrinome, ar atsako kintamasis „mpg“ yra pasiskirstytas pagal Gama arba Atvirkštinį Gauso skirstinį.



Pav 1 Atsako kintamojo histograma

Nubrėžus atsako kintamojo histogramą, Gama skirstinio su geriausiai parinktais scale ir shape parametrais tankio funkciją (Pav 1 pavaizduota raudonai) ir Atvirkštinio Gauso skirstinio su geriausiai parinktais miu ir lambda parametrais tankio funkciją (Pav 1 pavaizduota mėlinai), įsitikinome, jog atsako kintamasis išties yra pasiskirstytas pagal Gama arba Atvirkštinį Gauso skirstinį. Kolmogorovo-Smirnovo testo p-reikšmė gavosi lygi 0.1476 su Gama skirstiniu ir 0.0808 su Atvirkštiniu Gauso skirstiniu, kas yra daugiau už $\alpha = 0.05$. Reiškia, galime taikyti Gama arba Atvirkštinė Gauso regresiją.

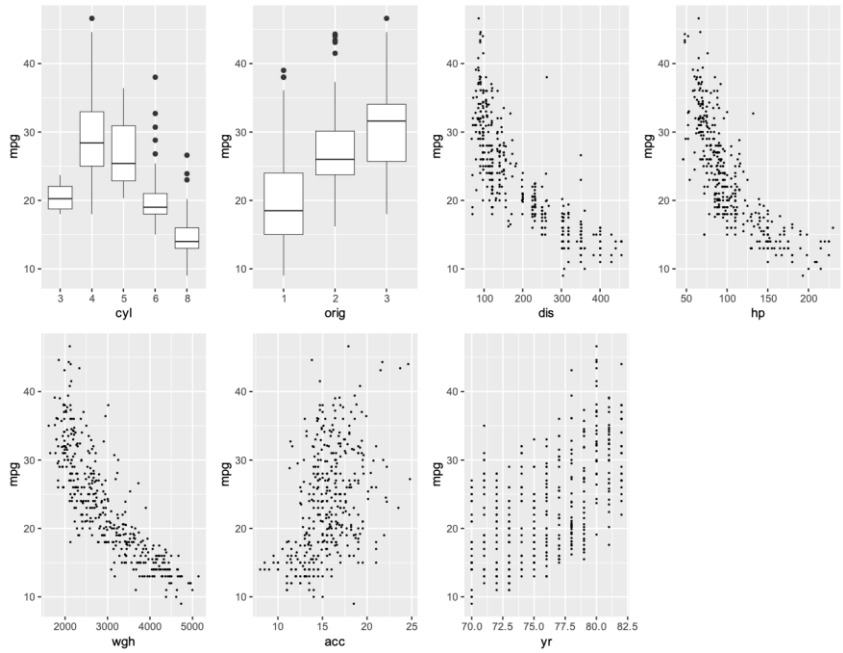
```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dt$mpg
D = 0.057656, p-value = 0.1476
alternative hypothesis: two-sided
```

Pav 2 Kolmogorovo-Smirnovo testo rezultatai Gama skirstiniui

```
    Asymptotic one-sample Kolmogorov-Smirnov test

data:  dt$mpg
D = 0.063975, p-value = 0.0808
alternative hypothesis: two-sided
```

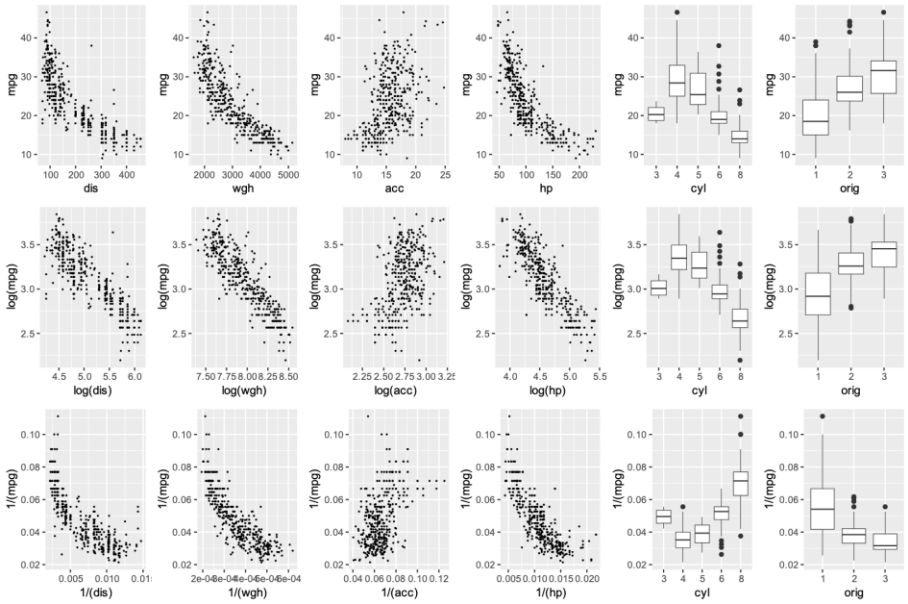
Pav 3 Kolmogorovo-Smirnovo testo rezultatai Atvirkštiniam Gauso skirstiniui

Ištyrus atsako kintamojo pasiskirstymą, nagrinėjome atsako kintamojo priklausomybes nuo kintamųjų. Iškarto atsisakėme nagrinėti automobilio markės kategorinį kintamąjį, nes jis turėjo virš 300 unikalių kategorijų. Kintamiesiems „cylinders“ ir „origin“ nubrėžėme stačiakampes diagramas, o kitiems – sklaidos. Išanalizavus grafikus, atsisakėme kintamojo „yr“, nurodančio automobilio pagaminimo metus, dėl didelio triukšmo.



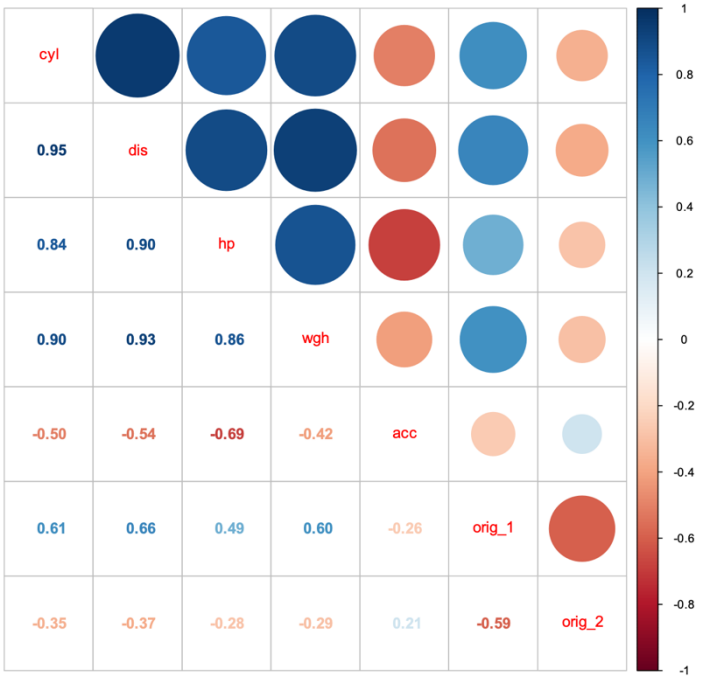
Pav 4 Atsako priklausomybė nuo kintamųjų

Kitus kintamuosius nusprendėme nagrinėti toliau ir papildomai jiems nubrėžėme atsako kintamojo priklausomybes nuo logaritmuotų ir apverstų duomenų.



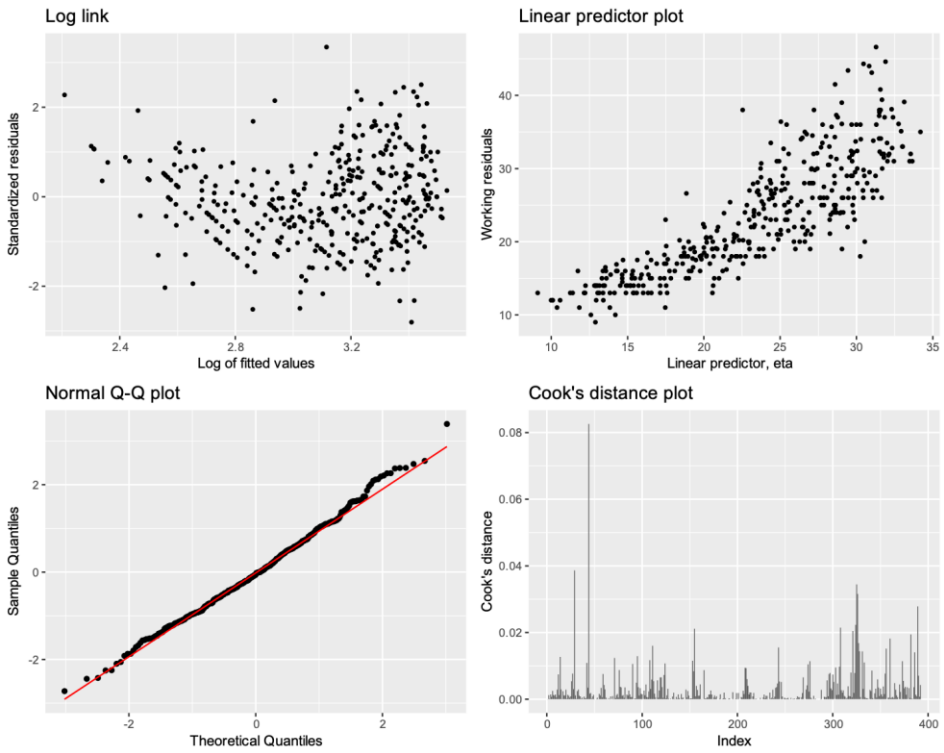
Pav 5 Papildomi grafikai

Gauti grafikai parodo kintamųjų naudojimo tinkamumą Gama ir Atvirkštinei Gauso regresijoms. Toliau tikrinome, ar tarp kintamųjų nėra reikšmingo multikolinearumo. Kadangi negalėjome apskaičiuoti koreliacijos tarp trijų kategorijų kintamojo „origin“ ir skaitinių kintamųjų, sukūrėme 2 naujus dichotominius kintamuosius orig_1 ir orig_2, kur orig_1 = 1, kai origin = 1, orig_1 = 0 kitais atvejais, o orig_2 = 1, kai orig_2 = 2 ir orig_2 = 0 kitais atvėjais. Tokiu būdu vienareikšmiškai pavirtome kintamąjį origin į du skaitinius kintamuosius. Toliau nubrėžėme koreliacijų matricą iš skaitinių kintamųjų ir naujų kategorinių indikatorių. Iš koreliacijų matricos gavome, jog tarp kintamųjų „cylinders“, „displacement“, „horsepower“ ir „weight“ paporinės koreliacijos yra nemažesnės už 0.84, kas reiškia, jog tarp šių kintamųjų yra didelis multikolinearumas. Todėl iš jų palikome tik vieną kintamąjį – „weight“, nes jis turėjo mažiausią absoliučią koreliaciją su kintamuoju „acceleration“ (0.42, kas yra tiknama).



Pav 6 Kintamųjų koreliacijų lentelė

Taigi, išmetus koreliuojančius kintamuosius, sukūrėme pradinį Gama regresijos modelį su „identity“ jungties funkcija ir regresoriais „weight“, „acceleration“ ir „origin“, nes jie tarpusavyje yra mažai koreliuoti. Toliau atlikome išimčių analizę. Iš standartizuotų liekanų, bei Cook‘o nuotolių grafikų priėjome išvadą, jog išskirčių nėra.



Pav 7 Liekanų grafikai

Išrinkus regresorius, bandėme surasti tiksliausią regresijos modelį. Iš viso lyginome 6 modelius: Gamma ir Atvirkštinės Gauso regresijos su skirtingomis jungties funkcijomis („identity“, „log“ ir „inverse“). Skirtingų modelių tikslumus vertinome, naudojant cross-validation metodą. Pirmiausiai duomenys padalinome į 10 imčių vienodo ilgio. Toliau iteruodami, ėmėme vieną iš imčių kaip testinę, o iš kitų formavome apmokymo imtį. Tada su atitinkama apmokymo imtimi kūrėme regresinį modelį. Sukurto modelio tikslumą vertinome su AIC ir kvadratinę paklaidą, kurią skaičiavome padavę modeliui testinius duomenis. Patikrinus modelį 10 kartų su skirtingomis apmokymo ir testavimo aibėmis, skaičiavome kvadratinių paklaidų vidurkį ir AIC.

names	AIC	MSE
Gamma log	2122.93	17.6285
Gamma identity	2159.62	18.2990
Gamma inverse	2241.54	25.2885
Inverse Gaussian log	2110.83	17.7357
Inverse Gaussian identity	2157.23	18.7149
Inverse Gaussian inverse	2227.44	31.1957

Pav 8 Regresijos modelių tikslumų lentelė

Patikrinus visus 6 modelius, nubrėžėme tikslumų lentelė. Iš jos matome, jog geriausias modelis pagal AIC (2110.83) yra Atvirkštinės Gauso regresijos modelis su logaritmine jungties funkcija. Geriausias modelis pagal vidutinę kvadratinę paklaidą gavosi Gama logaritminis regresijos modelis su MSE = 17.6285.

```
Call:
glm(formula = mpg ~ orig + log(wgh) + log(acc), family = inverse.gaussian(link = "log"),
    data = dt)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.27276    0.39063  26.298  < 2e-16 ***
orig2        0.02204    0.02678   0.823   0.4109
orig3        0.04873    0.02827   1.724   0.0855 .
log(wgh)    -0.98043    0.03981 -24.626  < 2e-16 ***
log(acc)     0.23065    0.04760   4.846 1.83e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for inverse.gaussian family taken to be 0.001173433)

    Null deviance: 2.02706  on 391  degrees of freedom
Residual deviance: 0.44602  on 387  degrees of freedom
AIC: 2110.8

Number of Fisher Scoring iterations: 5
```

Pav 9 Atvirkštinės Gauso regresijos modelio su logaritminė jungties funkcija santrauka

Galiausiai atsispausdinę dviejų geriausių modelių santrauką, sužinojome, jog Atvirkštinės Gauso regresijos modelis su logaritmine jungties funkcija turi žymiai mažesnį liekanų nuokrypį = 0.44602, kai Gama modelis su logaritminė jungties funkcija turi liekanų nuokrypį lygu 10.150.

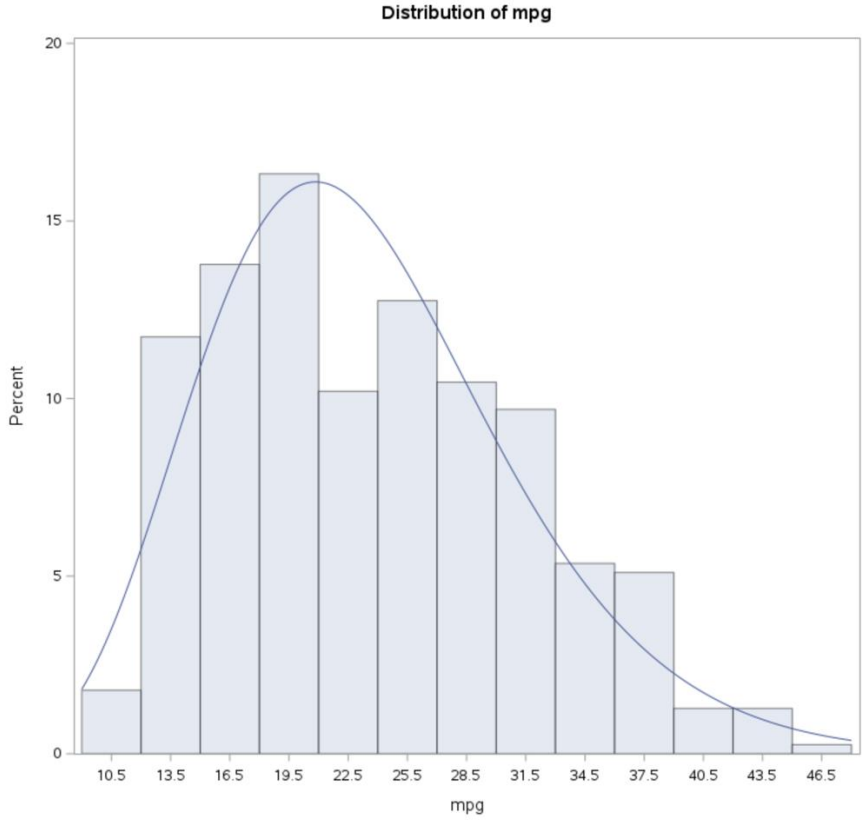
```
Call:
glm(formula = mpg ~ orig + log(wgh) + log(acc), family = Gamma(link = "log"),
    data = dt)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.11691    0.39837  25.396  < 2e-16 ***
orig2        0.01654    0.02558   0.647   0.5182
orig3        0.04579    0.02615   1.751   0.0808 .
log(wgh)    -0.96301    0.04060 -23.721  < 2e-16 ***
log(acc)     0.23722    0.05102   4.650 4.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.02704548)

    Null deviance: 44.205  on 391  degrees of freedom
Residual deviance: 10.150  on 387  degrees of freedom
AIC: 2122.9

Number of Fisher Scoring iterations: 4
```

Pav 10 Gama regresijos modelio su logaritminė jungties funkcija santrauka

Taigi suradome geriausiai tinkantį regresijos modelį, skirtą automobilio kuro sąnaudoms spėti, žinant jo kilmės regioną, svorį, bei 0-100 km/h įsibėgėjimo laiką. Toks modelis yra Atvirkštinės Gauso regresijos modelis su logaritminė jungties funkcija. Šio modelio AIC =2110.80, MSE = 17.6285 ir liekanų nuokrypis = 0.44602.

SPRENDIMAS SU SAS

Nuskaičius duomenis, atlikome pirminę duomenų analizę: panaikinome nulines stebinių reikšmes, bei, pasinaudojus Kolmogorovo-Smirnovo testu, patikrinome, ar atsako kintamasis „mpg“ yra pasiskirstytas pagal Gama skirstinį.



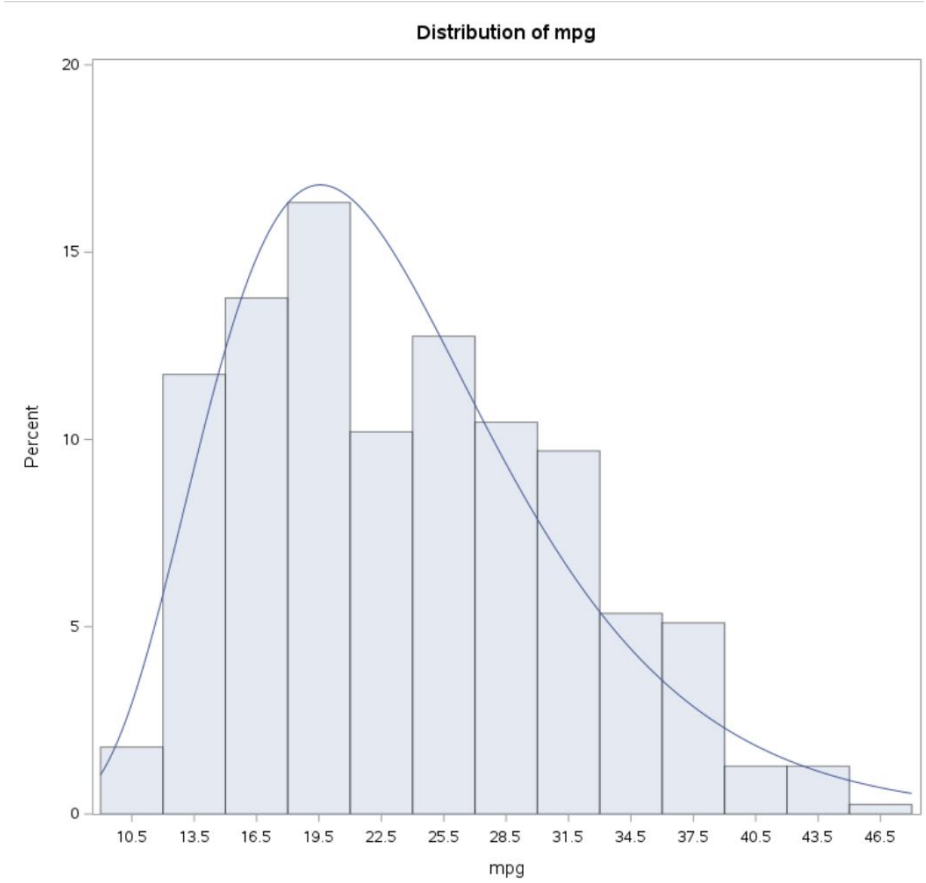
Pav 11 Atsako kintamojo histograma

Nubrėžus atsako kintamojo histogramą ir Gama skirstinio su geriausiai parinktais scale ir shape parametrais tankio funkciją, įsitikinome, jog atsako kintamasis išties yra pasiskirstytas pagal Gama skirstinį. Kolmogorovo-Smirnovo testo p-reikšmė gavosi lygi 0.142, kas yra daugiau už $\alpha = 0.05$. Reiškia, galime taikyti Gama regresiją.

Goodness-of-Fit Tests for Gamma Distribution				
Test	**Statistic**		**p Value**	
Kolmogorov-Smirnov	**D**	0.05777349	**Pr > D**	0.142
Cramer-von Mises	**W-Sq**	0.31484414	**Pr > W-Sq**	0.126
Anderson-Darling	**A-Sq**	2.00637309	**Pr > A-Sq**	0.093

Pav 12 Skirstinių lygybės testų rezultatai

Pasinaudojus Kolmogorovo-Smirnovo testu, patikrinome, ar atsako kintamasis „mpg“ yra pasiskirstytas pagal Atvirkštinį Gauso skirstinį.



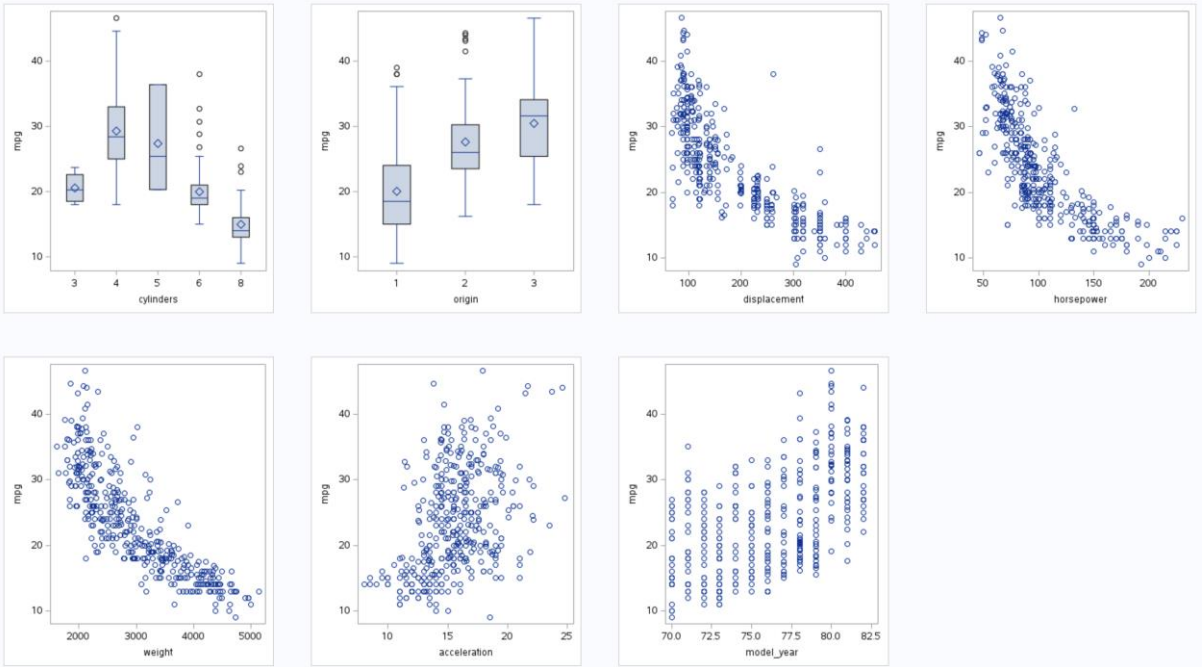
Pav 13 Atsako kintamojo histograma

Nubrėžus atsako kintamojo histogramą ir Atvirkštinio Gauso skirstinio su geriausiai parinktais mu ir lambda parametrais tankio funkciją, įsitikinome, jog atsako kintamasis išties yra pasiskirstytas pagal Atvirkštinį Gauso skirstinį. Kolmogorovo-Smirnovo testo p-reikšmė gavosi lygi 0.062, kas yra daugiau už $\alpha = 0.05$. Reiškia, galime taikyti Atvirkštinę Gauso regresiją.

Goodness-of-Fit Tests for Inverse Gaussian Distribution				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.06397536	Pr > D	0.062
Cramer-von Mises	W-Sq	0.33772024	Pr > W-Sq	0.106
Anderson-Darling	A-Sq	2.11437623	Pr > A-Sq	0.074

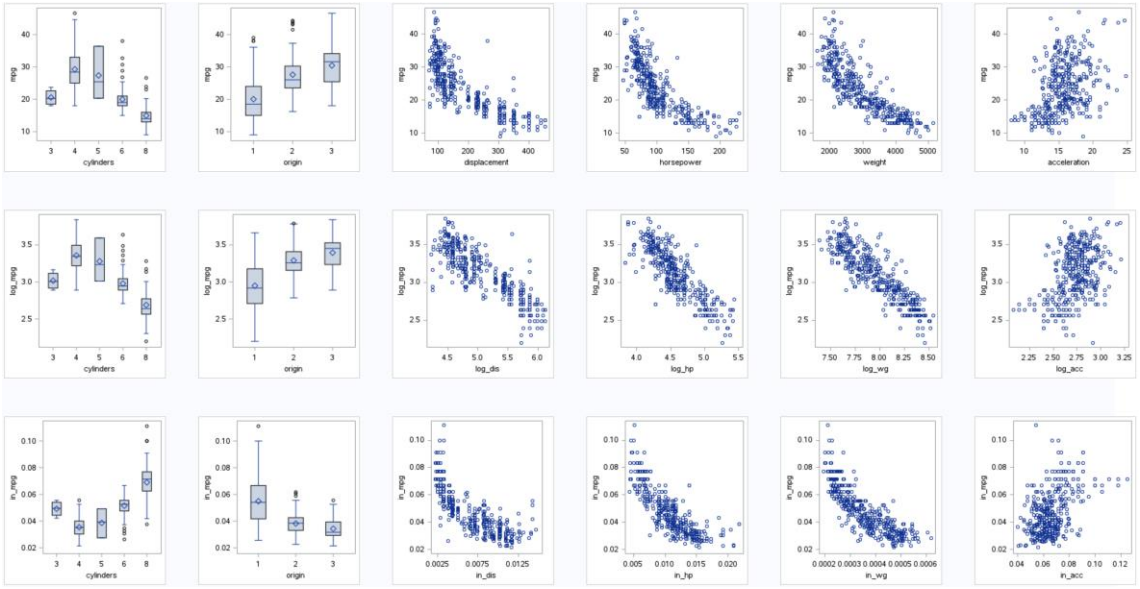
Pav 14 Skirstinių lygybės testų rezultatai

Ištyrus atsako kintamojo pasiskirstymą, nagrinėjome atsako kintamojo priklausomybes nuo kintamųjų. Iškarto atsisakėme nagrinėti automobilio markės kategorinį kintamąjį, nes jis turėjo virš 300 unikalių kategorijų. Kintamiesiems „cylinders“ ir „origin“ nubrėžėme stačiakampes diagramas, o kitiems – sklaidos. Išanalizavus grafikus, atsisakėme kintamojo „model_year“, nurodančio automobilio pagaminimo metus, dėl didelio triukšmo.



Pav 15 Atsako priklausomybė nuo kintamųjų

Kitus kintamuosius nusprendėme nagrinėti toliau ir papildomai jiems nubrėžėme atsako kintamojo priklausomybes nuo logaritmuotų ir apverstų duomenų.



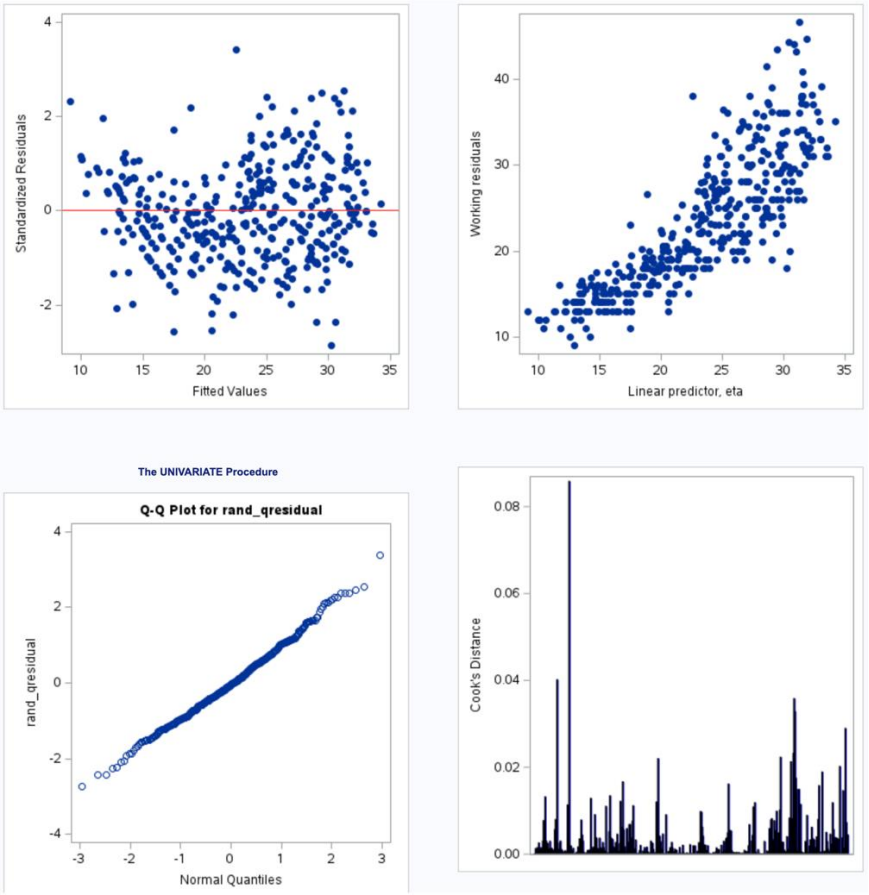
Pav 16 Papildomi grafikai

Gauti grafikai parodo kintamųjų naudojimo tinkamumą Gama ir Atvirkštinei Gauso regresijoms. Toliau tikrinome, ar tarp kintamųjų nėra reikšmingo multikolinearumo. Kadangi negalėjome apskaičiuoti koreliacijos tarp trijų kategorijų kintamojo „origin“ ir skaitinių kintamųjų, sukūrėme 2 naujus dichotominius kintamuosius orig_1 ir orig_2, kur orig_1 = 1, kai origin = 1, orig_1 = 0 kitais atvejais, o orig_2 = 1, kai orig_2 = 2 ir orig_2 = 0 kitais atvėjais. Tokiu būdu vienareikšmiškai pavirtome kintamąjį origin į du skaitinius kintamuosius. Toliau nubrėžėme koreliacijų matricą iš skaitinių kintamųjų ir naujų kategorinių indikatorių. Iš koreliacijų matricos gavome, jog tarp kintamųjų „cylinders“, „displacement“, „horsepower“ ir „weight“ paporinės koreliacijos yra nemažesnės už 0.84298, kas reiškia, jog tarp šių kintamųjų yra didelis multikolinearumas. Todėl iš jų palikome tik vieną kintamąjį – „weight“, nes jis turėjo mažiausią absoliučią koreliaciją su kintamuoju „acceleration“ (0.41684, kas yra tiknama).

Pearson Correlation Coefficients, N = 392 Prob > |r| under H0: Rho=0							
	orig_1	orig_2	cylinders	displacement	horsepower	weight	acceleration
orig_1	1.00000	-0.59143 <.0001	0.61049 <.0001	0.65594 <.0001	0.48962 <.0001	0.60098 <.0001	-0.25822 <.0001
orig_2	-0.59143 <.0001	1.00000	-0.35232 <.0001	-0.37163 <.0001	-0.28495 <.0001	-0.29384 <.0001	0.20830 <.0001
cylinders	0.61049 <.0001	-0.35232 <.0001	1.00000	0.95082 <.0001	0.84298 <.0001	0.89753 <.0001	-0.50468 <.0001
displacement	0.65594 <.0001	-0.37163 <.0001	0.95082 <.0001	1.00000	0.89726 <.0001	0.93299 <.0001	-0.54380 <.0001
horsepower	0.48962 <.0001	-0.28495 <.0001	0.84298 <.0001	0.89726 <.0001	1.00000	0.86454 <.0001	-0.68920 <.0001
weight	0.60098 <.0001	-0.29384 <.0001	0.89753 <.0001	0.93299 <.0001	0.86454 <.0001	1.00000	-0.41684 <.0001
acceleration	-0.25822 <.0001	0.20830 <.0001	-0.50468 <.0001	-0.54380 <.0001	-0.68920 <.0001	-0.41684 <.0001	1.00000

Pav 17 Kintamųjų koreliacijų lentelė

Taigi, išmetus koreliuojančius kintamuosius, modelio regresoriais parinkome kintamuosius „weight“, „acceleration“ ir „origin“, kurie tarpusavyje yra mažai koreliuoti. Toliau atlikome išimčių analizę. Iš standartizuotų liekanų, bei Cook‘o nuotolių grafikų priėjome išvadą, jog išskirčių nėra.



Pav 18 Liekanų grafikai

Išrinkus regresorius, bandėme surasti tiksliausią regresijos modelį. Kadangi SAS procedūra GENMOD nepalaiko atvirkštinės jungties funkcijos, teko lyginti tik 4 modelius: Gamma ir Atvirkštinės Gauso regresijos su „identity“ ir „log“ jungties funkcijomis. Skirtingų modelių tikslumus vertinome, naudojant cross-validation metodą. Pirmiausiai duomenys padalinome į 10 imčių vienodo ilgio. Toliau iteruodami, ėmėme vieną iš imčių kaip testinę, o iš kitų formavome apmokymo imtį. Tada su atitinkama apmokymo imtimi kūrėme regresinį modelį. Sukurto modelio tikslumą vertinome su AIC ir kvadratinę paklaidą, kurią skaičiavome padavę modeliui testinius duomenis. Patikrinus modelį 10 kartų su skirtingomis apmokymo ir testavimo aibėmis, skaičiavome kvadratinių paklaidų vidurkį ir AIC.

	NAME	AIC	MSE
1	Gamma_id	2159.6129	18.310805854
2	Gamma_lo	2122.9299	17.642072739
3	iGaus_id	2157.23	18.726590722
4	iGaus_lo	2110.8295	17.751191848

Pav 19 Regresijos modelių tikslumų lentelė

Patikrinus visus 6 modelius, nubrėžėme tikslumų lentelė. Iš jos matome, jog geriausias modelis pagal AIC (2110.8295) yra Atvirkštinės Gauso regresijos modelis su logaritmine jungties funkcija. Geriausias modelis pagal vidutinę kvadratinę paklaidą gavosi Gama logaritminis regresijos modelis su MSE = 17.6420

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	387	0.4460	0.0012
Scaled Deviance	387	392.0000	1.0129
Pearson Chi-Square	387	0.4541	0.0012
Scaled Pearson X2	387	399.1211	1.0313
Log Likelihood		-1049.4147	
Full Log Likelihood		-1049.4147	
AIC (smaller is better)		2110.8295	
AICC (smaller is better)		2111.0477	
BIC (smaller is better)		2134.6571	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	10.3214	0.3645	9.6069	11.0359	801.65	<.0001
origin_cat	1	1	-0.0487	0.0274	-0.1024	0.0049	3.17	0.0751
origin_cat	2	1	-0.0267	0.0304	-0.0862	0.0329	0.77	0.3798
origin_cat	3	0	0.0000	0.0000	0.0000	0.0000	.	.
log_wg		1	-0.9804	0.0386	-1.0561	-0.9048	645.69	<.0001
log_acc		1	0.2307	0.0462	0.1402	0.3211	24.96	<.0001
Scale		1	0.0337	0.0012	0.0315	0.0362		

Pav 20 Atvirkštinės Gauso regresijos modelio su logaritminė jungties funkcija santrauka

Galiausiai atsispausdinę dviejų geriausių modelių santrauką, sužinojome, jog Atvirkštinės Gauso regresijos modelis su logaritmine jungties funkcija turi žymiai mažesnį liekanų nuokrypį = 0.4460, kai Gama modelis su logaritminė jungties funkcija turi liekanų nuokrypį lygu 10.1500.

Taigi suradome geriausiai tinkantį regresijos modelį, skirtą automobilio kuro sąnaudoms spėti, žinant jo kilmės regioną, svorį, bei 0-100 km/h įsibėgėjimo laiką. Toks modelis yra Atvirkštinės Gauso regresijos modelis su logaritminė jungties funkcija. Šio modelio AIC =2110.8295, MSE = 17.6420 ir liekanų nuokrypis = 0.4460.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	387	10.1500	0.0262
Scaled Deviance	387	393.6843	1.0173
Pearson Chi-Square	387	10.4666	0.0270
Scaled Pearson X2	387	405.9626	1.0490
Log Likelihood		-1055.4650	
Full Log Likelihood		-1055.4650	
AIC (smaller is better)		2122.9299	
AICC (smaller is better)		2123.1481	
BIC (smaller is better)		2146.7575	

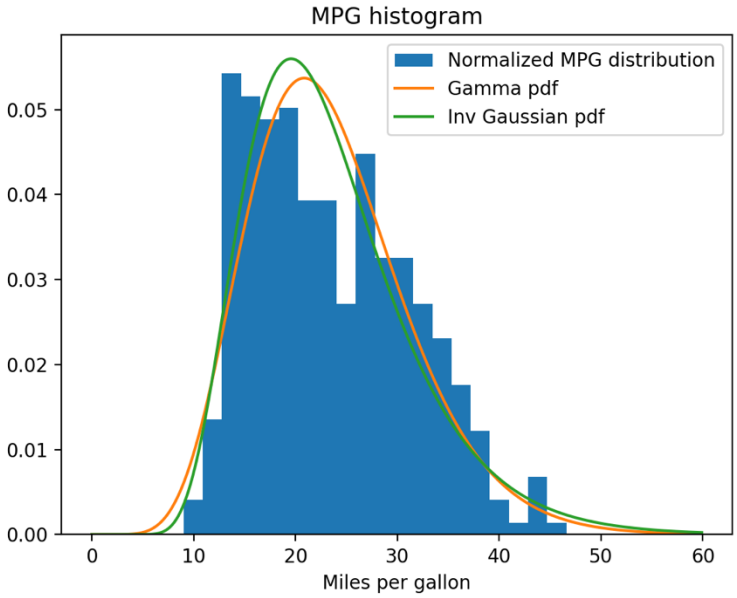
Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	10.1627	0.3713	9.4350	10.8905	749.06	<.0001
origin_cat	1	1	-0.0458	0.0253	-0.0953	0.0038	3.28	0.0702
origin_cat	2	1	-0.0292	0.0269	-0.0821	0.0236	1.18	0.2778
origin_cat	3	0	0.0000	0.0000	0.0000	0.0000	.	.
log_wg		1	-0.9630	0.0392	-1.0399	-0.8861	602.49	<.0001
log_acc		1	0.2372	0.0493	0.1405	0.3339	23.13	<.0001
Scale		1	38.7865	2.7586	33.7396	44.5883		

Pav 21 Gama regresijos modelio su logaritminė jungties funkcija santrauka

SPRENDIMAS SU PYTHON

Nuskaičius duomenis, atlikome pirminę duomenų analizę: panaikinome nulines stebinių reikšmes, bei, pasinaudojus Kolmogorovo-Smirnovo testu, patikrinome, ar atsako kintamasis „mpg“ yra pasiskirstytas pagal Gama arba Atvirkštinį Gauso skirstinį.



Pav 22 Atsako kintamojo histograma

Nubrėžus atsako kintamojo histogramą, Gama skirstinio su geriausiai parinktais scale ir shape parametrais tankio funkciją (Pav 22 pavaizduota raudonai) ir Atvirkštinio Gauso skirstinio su geriausiai parinktais miu ir lambda parametrais tankio funkciją (Pav 22 pavaizduota mėlinai), įsitikinome, jog atsako kintamasis išties yra pasiskirstytas pagal Gama arba Atvirkštinį Gauso skirstinį. Kolmogorovo-Smirnovo testo p-reikšmė gavosi lygi 0.1373 su Gama skirstiniu ir 0.0773 su Atvirkštiniu Gauso skirstiniu, kas yra daugiau už $\alpha = 0.05$. Reiškia, galime taikyti Gama arba Atvirkštinė Gauso regresiją.

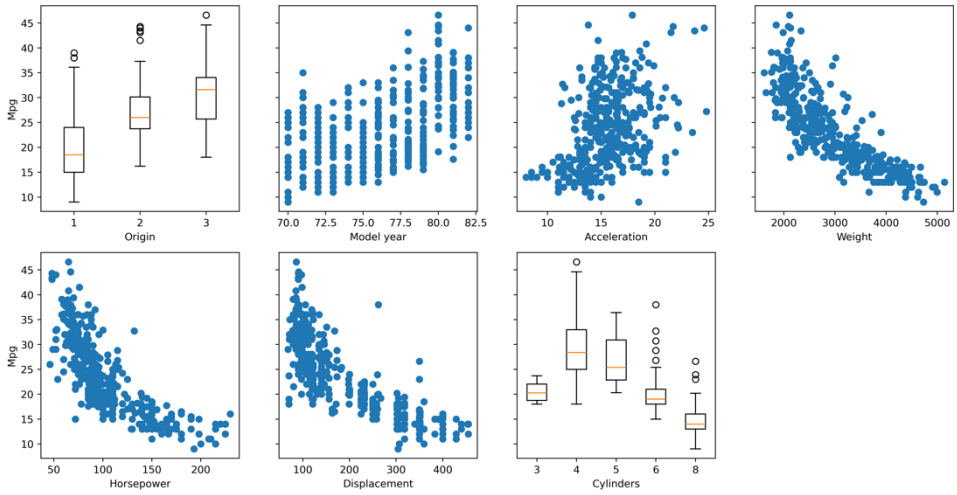
```
shape = 9.046845191319962,scale = 2.5916126419232013
KstestResult(statistic=0.05801499380561048, pvalue=0.13736761128520036, statistic_location=18.1, statistic_sign=1)
```

Pav 23 Kolmogorovo-Smirnovo testo rezultatai Gama skirstiniui

```
niu = 23.4459184099745,lambda = 193.383166845373
KstestResult(statistic=0.06397542559837899, pvalue=0.07734333345285094, statistic_location=26.0, statistic_sign=-1)
```

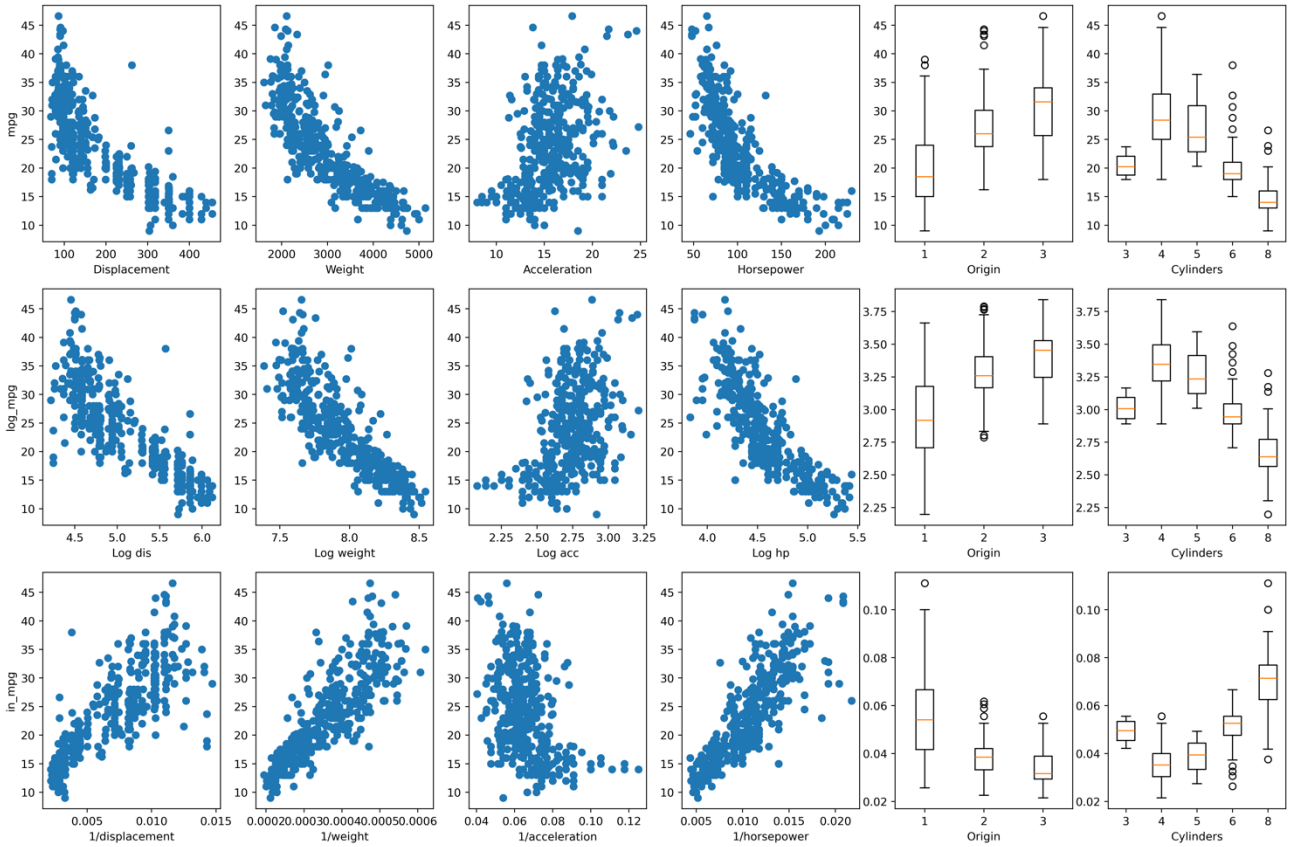
Pav 24 Kolmogorovo-Smirnovo testo rezultatai Atvirkštiniam Gauso skirstiniui

Ištyrus atsako kintamojo pasiskirstymą, nagrinėjome atsako kintamojo priklausomybes nuo kintamųjų. Iškarto atsisakėme nagrinėti automobilio markės kategorinį kintamąjį, nes jis turėjo virš 300 unikalių kategorijų. Kintamiesiems „cylinders“ ir „origin“ nubrėžėme stačiakampes diagramas, o kitiems – sklaidos. Išanalizavus grafikus, atsisakėme kintamojo „yr“, nurodančio automobilio pagaminimo metus, dėl didelio triukšmo.



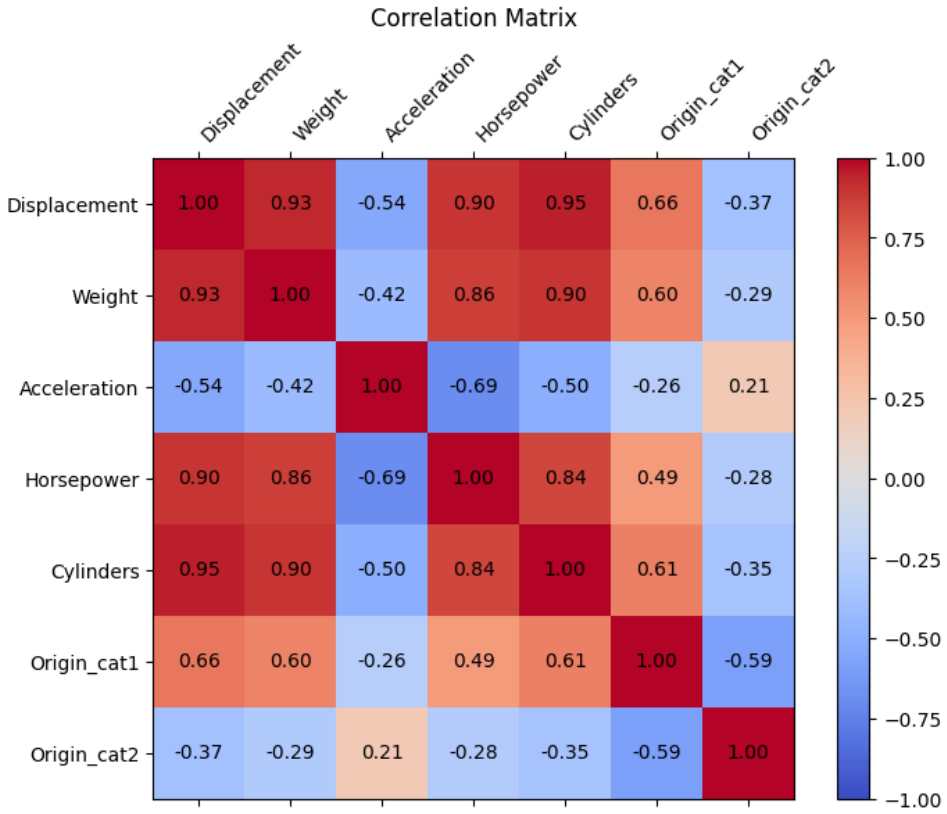
Pav 25 Atsako priklausomybė nuo kintamųjų

Kitus kintamuosius nusprendėme nagrinėti toliau ir papildomai jiems nubrėžėme atsako kintamojo priklausomybes nuo logaritmuotų ir apverstų duomenų.



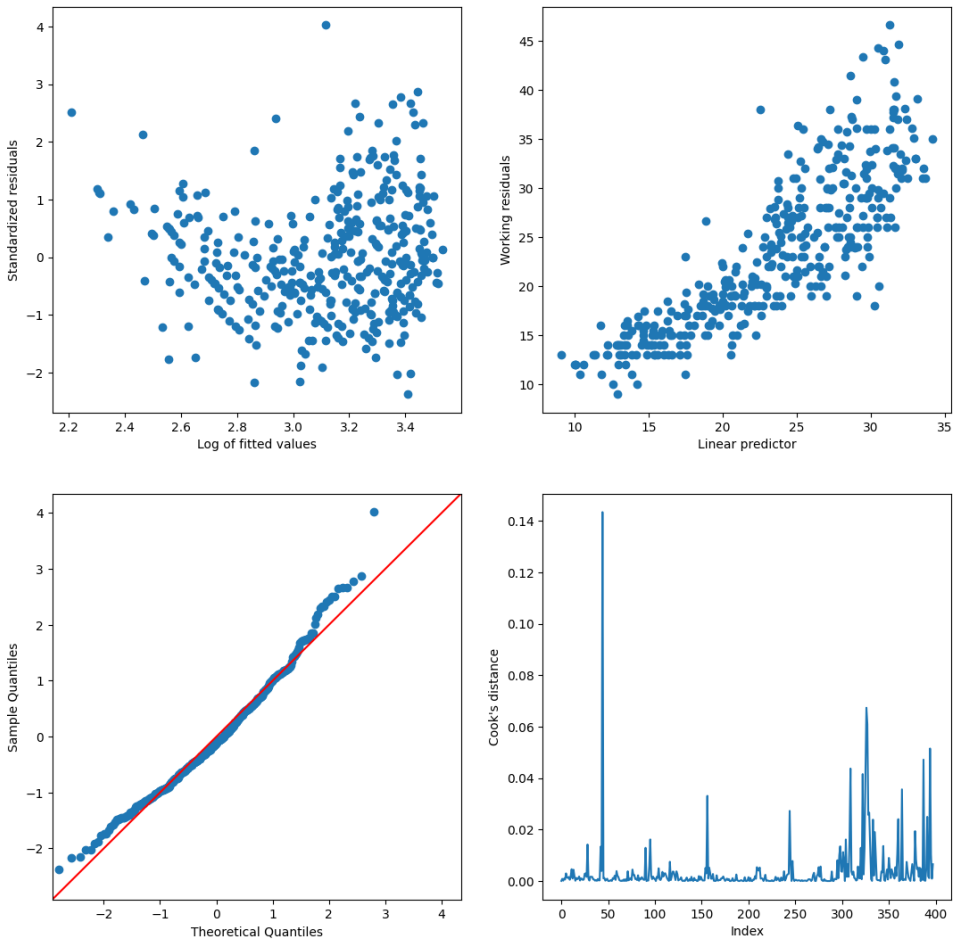
Pav 26 Papildomi grafikai

Gauti grafikai parodo kintamųjų naudojimo tinkamumą Gama ir Atvirkštinei Gauso regresijoms. Toliau tikrinome, ar tarp kintamųjų nėra reikšmingo multikolinearumo. Kadangi negalėjome apskaičiuoti koreliacijos tarp trijų kategorijų kintamojo „origin“ ir skaitinių kintamųjų, sukūrėme 2 naujus dichotominius kintamuosius origin_cat1 ir origin_cat2, kur origin_cat1 = 1, kai origin = 1, origin_cat1 = 0 kitais atvejais, o origin_cat2 = 1, kai origin = 2 ir origin_cat2 = 0 kitais atvėjais. Tokiu būdu vienareikšmiškai pavirtome kintamąjį origin į du skaitinius kintamuosius. Toliau nubrėžėme koreliacijų matricą iš skaitinių kintamųjų ir naujų kategorinių indikatorių. Iš koreliacijų matricos gavome, jog tarp kintamųjų „cylinders“, „displacement“, „horsepower“ ir „weight“ paporinės koreliacijos yra nemažesnės už 0.84, kas reiškia, jog tarp šių kintamųjų yra didelis multikolinearumas. Todėl iš jų palikome tik vieną kintamąjį – „weight“, nes jis turėjo mažiausią absoliučią koreliaciją su kintamuoju „acceleration“ (0.42, kas yra tiknama).



Pav 27 Kintamųjų koreliacijų lentelė

Taigi, išmetus koreliuojančius kintamuosius, sukūrėme pradinį Gama regresijos modelį su „identity“ jungties funkcija ir regresoriais „weight“, „acceleration“ ir „origin“, nes jie tarpusavyje yra mažai koreliuoti. Toliau atlikome išimčių analizę. Iš standartizuotų liekanų, bei Cook‘o nuotolių grafikų priėjome išvadą, jog išskirčių nėra.



Pav 28 Liekanų grafikai

Išrinkus regresorius, bandėme surasti tiksliausią regresijos modelį. Iš viso lyginome 6 modelius: Gamma ir Atvirkštinės Gauso regresijos su skirtingomis jungties funkcijomis („identity“, „log“ ir „inverse“). Skirtingų modelių tikslumus vertinome, naudojant 10 žingsnių cross-validation metodą. Python‘e, kryžminis patikrinimas buvo atliktas sklearn.model_selection.cross_validate pagalba. Sukurto modelio tikslumą kryžminio validavimo algoritme vertinome su kvadratinę paklaidą. Funkcija cross-validate atsitiktinai padalija paduodamus duomenis į 10 atsitiktinių, vienodo ilgio atkarpų. Patikrinus modelį 10 kartų su skirtingomis apmokymo ir testavimo aibėmis, kiekvienai iteracijai, algoritmas gražino kvadratinių paklaidų vidurkį (Mean Squared Error arba MSE). Paskaičiavus vidurkį visų MSE, gavome mūsų MSE įvertinį. Taip pat kiekvienam modeliui paskaičiavome AIC

	AIC	CV MSE
Gamma link=log	2121.375234	17.917912
Gamma link=identity	2157.906760	18.557387
Gamma link=inverse	2240.048840	25.577077
Inverse Gaussian link=log	2109.014004	18.018867
Inverse Gaussian link=identity	2155.332045	18.977253
Inverse Gaussian link=inverse	2225.684009	31.576425

Pav 29 Regresijos modelių tikslumų lentelė

Patikrinus visus 6 modelius, nubrėžėme tikslumų lentelė. Iš jos matome, jog geriausias modelis pagal AIC (2109.01) yra Atvirkštinės Gauso regresijos modelis su logaritmine jungties funkcija. Geriausias modelis pagal vidutinę kvadratinę paklaidą gavosi Gama regresijos modelis su logaritmine jungties funkciją su MSE = 17.92.

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                    mpg   No. Observations:                  392
Model:                            GLM   Df Residuals:                      387
Model Family:         InverseGaussian   Df Model:                            4
Link Function:                    Log   Scale:                       0.0011734
Method:                          IRLS   Log-Likelihood:                -1049.5
Date:                Sun, 24 Mar 2024   Deviance:                      0.44602
Time:                        12:29:08   Pearson chi2:                    0.454
No. Iterations:                    12   Pseudo R-squ. (CS):             0.9678
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.3214      0.378     27.330      0.000       9.581      11.062
log_weight    -0.9804      0.040    -24.626      0.000      -1.058      -0.902
log_acc        0.2307      0.048      4.846      0.000       0.137       0.324
origin_cat1   -0.0487      0.028     -1.724      0.085      -0.104       0.007
origin_cat2   -0.0267      0.031     -0.871      0.384      -0.087       0.033
==============================================================================
```

Pav 30 Atvirkštinės Gauso regresijos modelio su logaritminė jungties funkcija santrauka

Galiausiai atsispausdinę dviejų geriausių modelių santrauką, sužinojome, jog Atvirkštinės Gauso regresijos modelis su logaritmine jungties funkcija turi žymiai mažesnį liekanų nuokrypį = 0.44602, kai Gama modelis su logaritminė jungties funkcija turi liekanų nuokrypį lygu 10.150.

Generalized Linear Model Regression Results

Dep. Variable:	mpg	No. Observations:	392
Model:	GLM	Df Residuals:	387
Model Family:	Gamma	Df Model:	4
Link Function:	Log	Scale:	0.027045
Method:	IRLS	Log-Likelihood:	-1055.7
Date:	Sun, 24 Mar 2024	Deviance:	10.150
Time:	12:29:05	Pearson chi2:	10.5
No. Iterations:	10	Pseudo R-squ. (CS):	0.9597
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	10.1627	0.386	26.312	0.000	9.406	10.920
log_weight	-0.9630	0.041	-23.721	0.000	-1.043	-0.883
log_acc	0.2372	0.051	4.650	0.000	0.137	0.337
origin_cat1	-0.0458	0.026	-1.751	0.080	-0.097	0.005
origin_cat2	-0.0292	0.028	-1.063	0.288	-0.083	0.025

Pav 31 Gama regresijos modelio su logaritminė jungties funkcija santrauka

Taigi suradome geriausiai tinkantį regresijos modelį, skirtą automobilio kuro sąnaudoms spėti, žinant jo kilmės regioną, svorį, bei 0-100 km/h įsibėgėjimo laiką. Toks modelis yra Atvirkštinės Gauso regresijos modelis su logaritminė jungties funkcija. Šio modelio AIC = 2109.01, MSE = 18.02 ir liekanų nuokrypis = 0.44602.

IŠVADOS

Taigi, atlikome primynę duomenų analizę, patikrinome, ar tarp kintamųjų nėra multikolinearumo, atlikome išimčių analizę, išrinkome tinkamus regresorius, bei sudarėme 6 regresijos modelius (Gama ir Atvirkštinės Gauso su skirtingomis jungties funkcijomis), kuriuos palyginome, įvertinus jų tikslumus. Galiausiai suradome geriausiai tinkantį regresijos modelį, skirtą automobilio kuro sąnaudoms spėti, žinant jo kilmės regioną, svorį, bei 0-100 km/h įsibėgėjimo laiką. Toks modelis gavosi Atvirkštinės Gauso regresijos modelis su logaritminė jungties funkcija. Analizei atlikti naudojome SAS, Python ir R programavimo kalbas, tarp kuriu rezultatai sutapo. Taip atsitiko, kadangi modelio tikslumo analizei naudojome cross-validation metodą.

REGRESIJOS KOEFICIENTŲ INTERPRETACIJA

Kadangi geriausiai pasirodė Atvirkštinės Gauso regresijos modelis su logaritminė jungties funkcija, regresijos koeficientus interpretavome iš šio modelio.

Generalized Linear Model Regression Results

Dep. Variable:	mpg	No. Observations:	392
Model:	GLM	Df Residuals:	387
Model Family:	InverseGaussian	Df Model:	4
Link Function:	Log	Scale:	0.0011734
Method:	IRLS	Log-Likelihood:	-1049.5
Date:	Sun, 24 Mar 2024	Deviance:	0.44602
Time:	12:29:08	Pearson chi2:	0.454
No. Iterations:	12	Pseudo R-squ. (CS):	0.9678
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	10.3214	0.378	27.330	0.000	9.581	11.062
log_weight	-0.9804	0.040	-24.626	0.000	-1.058	-0.902
log_acc	0.2307	0.048	4.846	0.000	0.137	0.324
origin_cat1	-0.0487	0.028	-1.724	0.085	-0.104	0.007
origin_cat2	-0.0267	0.031	-0.871	0.384	-0.087	0.033

Pav 32 Atvirkštinės Gauso regresijos modelio su logaritminė jungties funkcija santrauka

Mūsų nagrinėjamas modelis yra:

$$EY = \exp(B_0 + B_1 \log(weight) + B_2 \log(acceleration) + B_3 origin1 + B_4 origin2)$$

$$EY = \exp(B_0) \cdot weight^{B_1} \cdot acceleration^{B_2} \cdot \exp(B_3 origin1) \cdot \exp(B_4 origin2)$$

Kur $B_0 = 10.3214$, $B_1 = -0.9804$, $B_2 = 0.2307$, $B_3 = -0.0487$ ir $B_4 = -0.0267$

Kintant kintamajam weight, EY keičiasi pagal šią funkciją:

$$(a \cdot weight)^{\beta_1} = a^{\beta_1} \cdot weight^{\beta_1}$$

Taigi kintamajam weight padidėjus a kartu, EY didėja a^{β_1} kartais.

Jei $a = 1.1, 1.1^{-0.9804} = 0.9107$, kas reiškia, jog padidinus kintamąjį weight 10 procentais, EY mažėja 8.93 procentais.

Kintant kintamajam acceleration , EY keičiasi pagal šią funkciją:

$$(a \cdot \text{acceleration})^{\beta_2} = a^{\beta_2} \cdot \text{acceleration}^{\beta_2}$$

Taigi kintamajam acceleration padidėjus a kartu, EY didėja a^2 kartais

Jei $a = 1.1, 1.1^{0.2307} = 1.0222$, kas reiškia, jog padidinus kintamąjį acceleration 10 procentais, EY didėja 2.22 procentais.

Toliau išnagrinėjome kaip kinta EY pereinant iš vienos origin kategorijos į kitą. Kadangi turėjome 3 kategorijas, sukūrėme perėjimų matricą.

A: $kur\ a_{ij}\ parodo\ kiek\ kartu\ pasikeite\ EY,\ pereinant\ kategorijai$ iš i į j

	1	2	3
1	-	$e^{\beta_4-\beta_3}$	$e^{-\beta_3}$
2	$e^{\beta_3-\beta_4}$	-	$e^{-\beta_4}$
3	e^{β_3}	e^{β_4}	-

Pav 33 EY santykiai pereinant iš atitinkamų kategorijų

	1	2	3
1	0	1.022244	1.049905
2	0.9782402	0	1.02706
3	0.9524668	0.9736533	0

Pav 34 EY santykiai su konkrečiais regresijos koeficientais

	1	2	3
1	0	2.2244 %	4.9905 %
2	-2.1750 %	0	2.7060 %
3	-4.7533 %	2.6347 %	0

Pav 35 EY kitimas procentais, priklausomai nuo perėjimo

Taigi, pereinant iš 1 kategorijos į 2, EY padidėja 2.2244 procentais. Kitos EY kitimo reikšmės pavaizduotos Pav 35

P.S.

Išnagrinėjus egzistuojančius pasiskirstymo tinkamumo testus, priėjome prie išvados, kad visi „sudėtingi“ testai iš esmės sudaryti iš dviejų dalių:

1. Numatomo skirstinio parametrų vertinimas iš duomenų.
2. „Paprasto“ pasiskirstymo tinkamumo testo pritaikymas, nurodant įvertintus parametrus.

Vienas iš tokių testų yra vadinamas Lilliefor‘o testu, kur norimo skirstinio parametrai yra vertinami didžiausio tikėtinumo metodu, o „paprastu“ testu yra pasirinktas Kolmogorovo-Smirnovo kriterijus. Kadangi prieš atliekant mūsų Kolmogorovo-Smirnovo testą, mes įvertinome skirstinio parametrus didžiausio tikėtinumo metodu, tai, iš tikrųjų, taikėme Lilliefor‘o testą, o ne Kolmogorovo-Smirnovo. Taigi, mes iš pat pradžių pritaikėme tinkamą „sudėtingą“ testą ir nieko keisti nereikia.