回帰分析（変数選択法）

東京国際大学 データサイエンス教育研究所 竹田 恒

2023-06-18

車の燃費データを使って回帰分析する。

1 データ

単位はオリジナル（出典参照）から馴染みのあるものに変換した。

出典：【UCI Machine Learning Repository】Auto MPG Data Set

説明変数	内容（単位）
燃費	燃費（km/L）
気筒数	気筒数（本）
排気量	排気量（cc）
馬力	馬力（hp）
車体重量	車体重量（kg）
加速性能	加速性能：時速97km（60mile）に到達する時間（秒）
製造年	製造年（年）
分類	アメ車，日本車，欧州車の3区分
車名	車名（英語）

```
d <- read.csv('https://stats.dip.jp/01_ds/data/car_mileage.csv')
rownames(d) <- paste0('No.', 1:nrow(d))

library(DT)
datatable(d)
```

Show 10 entries Search:

	燃費	気筒数	排気量	馬力	車体重量	加速性能	製造年	分類	車名
No.1	7.65	8	5031	130	1589	12	1970	アメ車	chevrolet chevelle malibu
No.2	6.38	8	5735	165	1675	11.5	1970	アメ車	buick skylark 320
No.3	7.65	8	5211	150	1559	11	1970	アメ車	plymouth satellite
No.4	6.8	8	4982	150	1557	12	1970	アメ車	amc rebel sst
No.5	7.23	8	4949	140	1564	10.5	1970	アメ車	ford torino
No.6	6.38	8	7030	198	1969	10	1970	アメ車	ford galaxie 500
No.7	5.95	8	7440	220	1975	9	1970	アメ車	chevrolet impala
No.8	5.95	8	7210	215	1956	8.5	1970	アメ車	plymouth fury iii
No.9	5.95	8	7456	225	2007	10	1970	アメ車	pontiac catalina
No.10	6.38	8	6391	190	1746	8.5	1970	アメ車	amc ambassador dpl

Showing 1 to 10 of 392 entries

Previous 1 2 3 4 5 … 40 Next

2 フィッティング

車名を除くすべての説明変数を使って回帰分析する。【注意】車名を入れると変数の数がデータサイズに近づき過学習する。

```
fit.full <- lm(燃費 ~ 気筒数 + 排気量 + 馬力 + 車体重量 + 加速性能 + 製造年 + 分類, data = d)
#summary(fit.full)
library(sjPlot)
tab_model(fit.full, show.stat = T, show.aic = T)
```

	燃費			
Predictors	Estimates	CI	Statistic	p
(Intercept)	-635.52	-721.28 – -549.77	-14.57	**<0.001**
気筒数	-0.21	-0.48 – 0.06	-1.52	0.128
排気量	0.00	0.00 – 0.00	3.14	**0.002**
馬力	-0.01	-0.02 – 0.00	-1.33	0.185
車体重量	-0.01	-0.01 – -0.01	-10.24	**<0.001**
加速性能	0.03	-0.05 – 0.12	0.81	0.421
製造年	0.33	0.29 – 0.37	15.01	**<0.001**
分類 [日本車]	1.21	0.75 – 1.67	5.16	**<0.001**
分類 [欧州車]	1.12	0.64 – 1.59	4.64	**<0.001**
Observations	392			
R^2 / R^2 adjusted	0.824 / 0.821			
AIC	1390.309			

3 変数を手動選択

有意でない偏回帰係数を持つ説明変数を除外

```
fit.manual <- lm(燃費 ~ 排気量 + 車体重量 + 製造年 + 分類, data = d)
#summary(fit.manual)
tab_model(fit.manual, show.stat = T, show.aic = T)
```

	燃費			
Predictors	Estimates	CI	Statistic	p
(Intercept)	-653.14	-737.14 – -569.15	-15.29	**<0.001**
排気量	0.00	-0.00 – 0.00	1.87	0.063
車体重量	-0.01	-0.01 – -0.01	-12.10	**<0.001**
製造年	0.34	0.30 – 0.38	15.71	**<0.001**
分類 [日本車]	1.04	0.59 – 1.48	4.59	**<0.001**
分類 [欧州車]	1.01	0.54 – 1.48	4.25	**<0.001**
Observations	392			
R^2 / R^2 adjusted	0.821 / 0.818			
AIC	1392.150			

4 ステップワイズ変数選択

変数間の相乗効果を表す**交互作用項**を追加（モデル式を2乗）する。 交互変数を追加すると組み合わせ数が多くなり，手動では変数選択が難しいため， ステップワイズ変数選択法を使用する。

【注意】ステップワイズ変数選択では有意でない偏回帰係数も選択される。 これはモデル選択の基準となるAICが予測誤差最小を評価する指標であり， 偏回帰係数の有意性の有無は考慮していないためである。

```
library(MASS)
fit.aic <- stepAIC(lm(燃費 ~ (気筒数 + 排気量 + 馬力 + 車体重量 + 加速性能 + 製造年 + 分類)^2, d
ata = d), trace = 0)
#summary(fit.aic)
tab_model(fit.aic, show.stat = T, show.aic = T)
```

	燃費			
Predictors	Estimates	CI	Statistic	p
(Intercept)	-24.78	-777.54 – 727.99	-0.06	0.948
気筒数	-1.02	-1.77 – -0.26	-2.65	**0.008**
排気量	-0.09	-0.20 – 0.02	-1.59	0.113
馬力	5.75	1.12 – 10.38	2.44	**0.015**
車体重量	-0.01	-0.01 – -0.01	-8.58	**<0.001**
加速性能	-51.71	-84.66 – -18.77	-3.09	**0.002**
製造年	0.03	-0.35 – 0.41	0.15	0.881
分類 [日本車]	-214.70	-421.34 – -8.06	-2.04	**0.042**
分類 [欧州車]	-232.77	-448.44 – -17.09	-2.12	**0.034**
気筒数 × 加速性能	0.07	0.02 – 0.11	2.89	**0.004**
排気量 × 車体重量	0.00	0.00 – 0.00	7.35	**<0.001**
排気量 × 製造年	0.00	-0.00 – 0.00	1.55	0.122
排気量 × 分類 [日本車]	0.00	0.00 – 0.00	3.78	**<0.001**
排気量 × 分類 [欧州車]	-0.00	-0.00 – 0.00	-0.90	0.369
馬力 × 製造年	-0.00	-0.01 – -0.00	-2.45	**0.015**
車体重量 × 分類 [日本車]	-0.00	-0.01 – -0.00	-2.46	**0.014**
車体重量 × 分類 [欧州車]	0.00	-0.00 – 0.00	0.48	0.634
加速性能 × 製造年	0.03	0.01 – 0.04	3.06	**0.002**
加速性能 × 分類 [日本車]	0.34	0.15 – 0.54	3.45	**0.001**
加速性能 × 分類 [欧州車]	0.44	0.29 – 0.59	5.87	**<0.001**
製造年 × 分類 [日本車]	0.11	0.00 – 0.21	2.00	**0.046**
製造年 × 分類 [欧州車]	0.11	0.01 – 0.22	2.06	**0.040**
Observations	392			
R^2 / R^2 adjusted	0.894 / 0.888			
AIC	1217.342			

5 モデル比較

```
tab_model(fit.full, fit.manual, fit.aic,
          show.stat = T, show.aic = T,
          dv.labels = c('車名除きフルモデル',
                        '手動選択モデル',
                        'ステップワイズ変数選択モデル'))
```

	車名除きフルモデル				手動選択モデル				ステップワイズ変数選択モデル			
Predictors	Estimates	CI	Statistic	p	Estimates	CI	Statistic	p	Estimates	CI	Statistic	p
(Intercept)	-635.52	-721.28 – -549.77	-14.57	**<0.001**	-653.14	-737.14 – -569.15	-15.29	**<0.001**	-24.78	-777.54 – 727.99	-0.06	0.948
気筒数	-0.21	-0.48 – 0.06	-1.52	0.128					-1.02	-1.77 – -0.26	-2.65	**0.008**
排気量	0.00	0.00 – 0.00	3.14	**0.002**	0.00	-0.00 – 0.00	1.87	0.063	-0.09	-0.20 – 0.02	-1.59	0.113
馬力	-0.01	-0.02 – 0.00	-1.33	0.185					5.75	1.12 – 10.38	2.44	**0.015**
車体重量	-0.01	-0.01 – -0.01	-10.24	**<0.001**	-0.01	-0.01 – -0.01	-12.10	**<0.001**	-0.01	-0.01 – -0.01	-8.58	**<0.001**
加速性能	0.03	-0.05 – 0.12	0.81	0.421					-51.71	-84.66 – -18.77	-3.09	**0.002**
製造年	0.33	0.29 – 0.37	15.01	**<0.001**	0.34	0.30 – 0.38	15.71	**<0.001**	0.03	-0.35 – 0.41	0.15	0.881
分類 [日本車]	1.21	0.75 – 1.67	5.16	**<0.001**	1.04	0.59 – 1.48	4.59	**<0.001**	-214.70	-421.34 – -8.06	-2.04	**0.042**
分類 [欧州車]	1.12	0.64 – 1.59	4.64	**<0.001**	1.01	0.54 – 1.48	4.25	**<0.001**	-232.77	-448.44 – -17.09	-2.12	**0.034**
気筒数 × 加速性能									0.07	0.02 – 0.11	2.89	**0.004**
排気量 × 車体重量									0.00	0.00 – 0.00	7.35	**<0.001**
排気量 × 製造年									0.00	-0.00 – 0.00	1.55	0.122
排気量 × 分類 [日本車]									0.00	0.00 – 0.00	3.78	**<0.001**
排気量 × 分類 [欧州車]									-0.00	-0.00 – 0.00	-0.90	0.369
馬力 × 製造年									-0.00	-0.01 – -0.00	-2.45	**0.015**
車体重量 × 分類 [日本車]									-0.00	-0.01 – -0.00	-2.46	**0.014**
車体重量 × 分類 [欧州車]									0.00	-0.00 – 0.00	0.48	0.634
加速性能 × 製造年									0.03	0.01 – 0.04	3.06	**0.002**
加速性能 × 分類 [日本車]									0.34	0.15 – 0.54	3.45	**0.001**
加速性能 × 分類 [欧州車]									0.44	0.29 – 0.59	5.87	**<0.001**
製造年 × 分類 [日本車]									0.11	0.00 – 0.21	2.00	**0.046**
製造年 × 分類 [欧州車]									0.11	0.01 – 0.22	2.06	**0.040**
Observations	392				392				392			
R^2 / R^2 adjusted	0.824 / 0.821				0.821 / 0.818				0.894 / 0.888			
AIC	1390.309				1392.150				1217.342			

> 自由度調整済み決定係数（R^2 adjusted）やAICを比べると， は**ステップワイズ変数選択モデル**，次いで**車名除きフルモデル**， 最後に**手動選択モデル**の順に良いモデルとなった。

6 演習課題

赤ワインの品質（quality）を予測する重回帰モデルとして， フルモデル，手動選択モデル，ステップワイズ変数選択モデルを作成し性能を評価せよ。
ステップワイズ変数選択では， 化学物質の組み合わせの妙を表現するための交互作用項を導入せよ。

データの内容は次のURLを参照のこと。
赤ワインデータの解析

```
d <- read.csv('https://stats.dip.jp/01_ds/data/winequality-red.csv')
datatable(d)
```

Show 10 entries Search:

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
3	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
8	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
9	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
10	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5

Showing 1 to 10 of 1,599 entries

Previous 1 2 3 4 5 … 160 Next