RAPPORT DU PROJET

Course: LSTAT 2120 Linear Model

Authors:	Mathias Dah Fienon	&	Gabin Kana Tsiguia

Prof: Christian Hafner

Assistant: Stefka Asenova

Academic year: 2021-2022

About the dataset and the objectives of this study:

Our dataset is a set of 400 different stores, each one described by:

- their total sales (in thousands of dollars)
- the budget invested in advertisement (in thousands of dollars also)
- the income levels of located region of the stores
- the population of those region
- their age
- the levels of education
- the price charged by our targeted stores
- the price charged by the competitor of the said stores
- and the quality of the shelving location for the car seats at each store.

variables which we can observe in the table of **Fig A1** in the appendix, that there are 8 numerical variables and 3 categorical variables.

The idea is to set a model based on the up said variables to predict the sales of the stores. First, we go on the hypothesis that variables like income, advertising, CompPrice, Price and age are more likely to predict the sales of the stores.

For this, we are going to start with a descriptive analysis of our variables, then, use some method to select the adequate variables to include in the model, verify the underlying hypotheses of the model such as linearity, multicollinearity, heteroskedasticity, etc. After, we will make some test on the coefficient estimate and some prediction to evaluate the accuracy of our model.

DESCRIPTIVE ANALYSIS:

La procédure MEANS

Variable	N	Moyenne	Ec-type	Erreur type	25ème ctl	Médiane	50ème ctl	Maximum	Minimum	Skewness	Kurtosis
Advertising	400	6.635	6.650	0.333	0.000	5.000	5.000	29.000	0.000	0.640	-0.545
Age	400	53.323	16.200	0.810	39.500	54.500	54.500	80.000	25.000	-0.077	-1.134
CompPrice	400	124.975	15.335	0.767	115.000	125.000	125.000	175.000	77.000	-0.043	0.042
Income	400	68.658	27.986	1.399	42.500	69.000	69.000	120.000	21.000	0.049	-1.085
Population	400	264.840	147.376	7.369	139.000	272.000	272.000	509.000	10.000	-0.051	-1.202
Price	400	115.795	23.677	1.184	100.000	117.000	117.000	191.000	24.000	-0.125	0.452
Sales	400	7.496	2.824	0.141	5.380	7.490	7.490	16.270	0.000	0.186	-0.081

Table 1: Summary statistics of the variables

The analysis of the budget invested by each of the stores, shows that most of the stores invested more than 5000$ in advertisement towards the clients. In average 6635$ is invested in advertisement with a standardized deviation around 6.65$.

In contrast with the sales, we see that in average, the sales are about 7496$, for a maximum sale of 16270$, where more than 50% of the stores sold for at least 7490$ of child car seats.

The analysis of the sales with respect to the shelving location of the car seats in the stores (Fig1), let us confirm an obvious intuition "More the product is well located in the stores, higher are the sales".

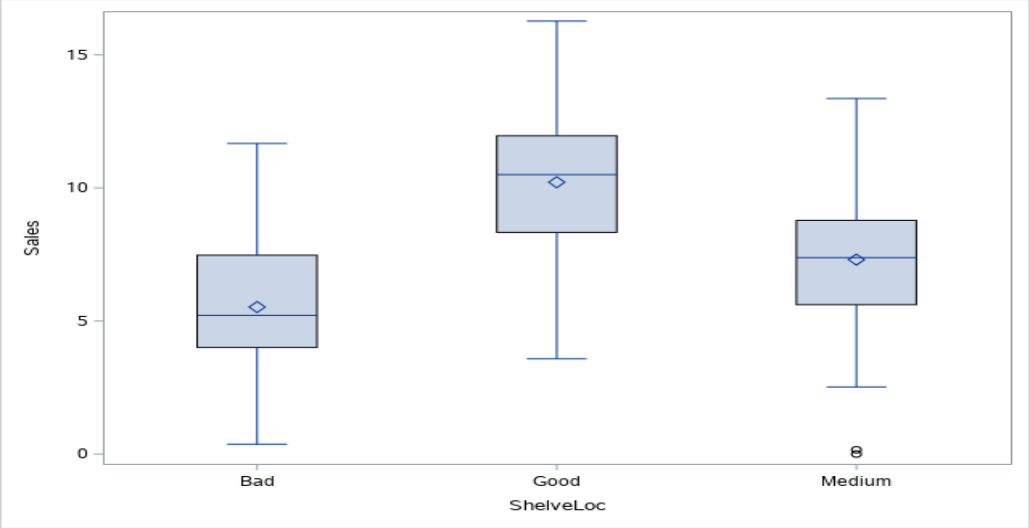


Fig 1: Boxplot Sales by Shelving Location

By analyzing in first sight the Sales type of region (Urban or not), we can say that the location of the stores, in urban area of not, has no incident on the sales. The average sales are almost the same in both cases, as shown in the following two tables

Urban	Méthode	Moyenne	IC à95% - Moyenne		Ec-type	Ec.-type de l'IC à95%	
No		7.5636	7.0520	8.0751	2.8058	2.4878	3.2179
Yes		7.4682	7.1357	7.8007	2.8362	2.6199	3.0918
Diff (1-2)	Pooled	0.0954	-0.5140	0.7048	2.8273	2.6438	3.0384
Diff (1-2)	Satterthwaite	0.0954	-0.5128	0.7036			

Méthode	Variances	DDL	Valeur du test t	Pr > |t|
Pooled	Egal	398	0.31	0.7585
Satterthwaite	Non égal	221.57	0.31	0.7576

Egalité des variances				
Méthode	DDL num.	DDL den.	Valeur F	Pr > F
Folded F	281	117	1.02	0.9072

Table 2: Comparison of means of sales by region (urban or not)

The cross-table analysis of region (US or No US) with the ShelveLoc, let us suggest that neither the stores is located in the US or not, the quality of the shelving location for the car seats is medium. The chi2 test of independency put out an independence between the located region of the stores and the quality of the shelving location of the car seats in the said store....

Statistiques pour la table de US par ShelveLoc

Statistique	DDL	Valeur	Prob
Khi-2	2	2.7397	0.2541
Test du rapport de vraisemblance	2	2.8024	0.2463
Khi-2 de Mantel-Haenszel	1	0.6308	0.4271
Coefficient Phi		0.0828	
Coefficient de contingence		0.0825	
V de Cramer		0.0828	

Taille de l'échantillon = 400

Fig 4: Independency test of US and ShelveLoc

And at last, let's compare the competitor pricing with the price indicated in each store. In that, the average price of the competitor is 124$, the tenth first price charged is 88.5$ and a modal price of

121$ where in the targeted stores the average price charged is 115$, the tenth first price charged is 54.5$ and the modal price about 120$. This show that our targeted stores are well placed to have the most sales. But that will be for another study.

Above all, the variable income has its word in this analysis because it is the in somehow an endogen variable because all decision in based on that. So, the analysis of the income of the population in the area where our targeted store is located, shows us that the average income in all those regions is 68657$ with a standard deviation of 27.98$. 50% of the population in the targeted region earn between 42500$ and 91000$. In add, the comparison between Urban area and none urban area doesn't show any glaring difference. The same remark is made about the US region or none US region.

Coefficienti di correlazione di Pearson, N = 400 Prob > |r| sotto H0: Rho=0								
	Sales	CompPrice	Income	Advertising	Population	Price	Age	Education
Sales	1.00000	0.06408 0.2009	0.15195 0.0023	0.26951 <.0001	0.05047 0.3140	-0.44495 <.0001	-0.23182 <.0001	-0.05196 0.2999
CompPrice	0.06408 0.2009	1.00000	-0.08065 0.1073	-0.02420 0.6294	-0.09471 0.0584	0.58485 <.0001	-0.10024 0.0451	0.02520 0.6154
Income	0.15195 0.0023	-0.08065 0.1073	1.00000	0.05899 0.2391	-0.00788 0.8752	-0.05670 0.2579	-0.00467 0.9258	-0.05686 0.2566
Advertising	0.26951 <.0001	-0.02420 0.6294	0.05899 0.2391	1.00000	0.26565 <.0001	0.04454 0.3743	-0.00456 0.9276	-0.03359 0.5029
Population	0.05047 0.3140	-0.09471 0.0584	-0.00788 0.8752	0.26565 <.0001	1.00000	-0.01214 0.8087	-0.04266 0.3948	-0.10638 0.0334
Price	-0.44495 <.0001	0.58485 <.0001	-0.05670 0.2579	0.04454 0.3743	-0.01214 0.8087	1.00000	-0.10218 0.0411	0.01175 0.8148
Age	-0.23182 <.0001	-0.10024 0.0451	-0.00467 0.9258	-0.00456 0.9276	-0.04266 0.3948	-0.10218 0.0411	1.00000	0.00649 0.8971
Education	-0.05196 0.2999	0.02520 0.6154	-0.05686 0.2566	-0.03359 0.5029	-0.10638 0.0334	0.01175 0.8148	0.00649 0.8971	1.00000

Table 2: correlation matrix

The correlation matrix let suggest a:

- negative correlation between Sales and Price: showing that if the price increase, but sales are likely to decrease.
-

VARIABLES SELECTION:

Prior procedures:

Now, let's jump onto the variable selection for our prediction model.

For that as know, we first have to render dummies the categorical variables of our dataset. Recall that ShelveLoc, Urban and US are the categorical variables, so render them dummy (and omitting the last modality of each variable).

So, in the general case, we specify this model:

$$\boldsymbol{sales} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * \boldsymbol{CompPrice} + \boldsymbol{\beta_2} * \boldsymbol{Income} + \boldsymbol{\beta_3} * \boldsymbol{Advertising} + \boldsymbol{\beta_4} * \boldsymbol{Population} + \boldsymbol{\beta_5} * \boldsymbol{Price} + \boldsymbol{\beta_6} * \boldsymbol{Age} + \boldsymbol{\beta_7} * \boldsymbol{Education} + \boldsymbol{\beta_8} * \boldsymbol{ShelveLoc_{Bad}} + \boldsymbol{\beta_9} * \boldsymbol{ShelveLoc_{Good}} + \boldsymbol{\beta_{10}} * \boldsymbol{Urban_{Yes}} + \boldsymbol{\beta_{11}} * \boldsymbol{US_{Yes}} + \boldsymbol{\epsilon}$$

After the specification of the general model, let's split the dataset into two parts: one to use for model selection and parameter estimation and the second one for prediction. For that we select randomly 100 observations to set apart for the prediction and 300 observations to train and select the model.

The 300 observations dataset is now split into two sets the train set (210 observations) and the test set (90 observations) respectively 70% and 30%.

Variable's selections

Firstly, we compute a regression model based on all the available variables in the dataset with the "stepwise method" with the default ***significance level to stay and to entry fixed at 0.15.***

Synthèse de Sélection Stepwise

Etape	Variable entrée	Variable supprimée	Nombre var. dans	R carré partiel	R carré du modèle	C(p)	Valeur F	Pr > F
1	ShelGood		1	0.3059	0.3059	797.909	91.66	<.0001
2	Price		2	0.1999	0.5058	510.734	83.75	<.0001
3	CompPrice		3	0.1324	0.6382	321.244	75.39	<.0001
4	Advertising		4	0.0686	0.7068	224.048	47.95	<.0001
5	Age		5	0.0686	0.7754	126.813	62.32	<.0001
6	ShelBad		6	0.0539	0.8293	50.8880	64.07	<.0001
7	Income		7	0.0321	0.8614	6.5214	46.71	<.0001

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	7	1415.70696	202.24385	179.28	<.0001
Erreur	202	227.87395	1.12809		
Total sommes corrigées	209	1643.58091			

Root MSE	1.06212	R carré	0.8614
Moyenne dépendante	7.54267	R car. ajust.	0.8566
Coeff Var	14.08143		

Table 3: stepwise model selection method ouput

Based on the stepwise method, the variables ***CompPrice, Income, Advertising, Price, Age, Shel_Bad*** and ***Shel_Good*** are the one selected with an **adjusted R2 = 0.8566** and a **mallow criterion (cp) = 6.5214**.

Using the mallow criterion (FigA2) and a trade-off of less variables, the model with less variables and the smaller cp is the one with the variable ***CompPrice Income Advertising Price Age ShelBad*** and ***ShelGood.*** Same as the stepwise has chosen.

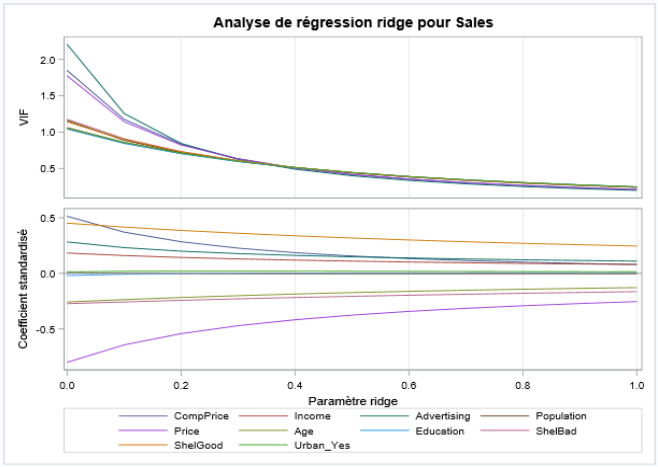
Also, as we can see in the table below, with the selected variable, there is no multicollinearity problem as all the variance inflation factor are under 10.

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Inflation de variance
Intercept	1	6.71616	0.73742	9.11	<.0001	0
CompPrice	1	0.09265	0.00631	14.69	<.0001	1.76843
Income	1	0.01906	0.00279	6.83	<.0001	1.05439
Advertising	1	0.11008	0.01176	9.36	<.0001	1.04453
Price	1	-0.09088	0.00391	-23.22	<.0001	1.73555
Age	1	-0.04706	0.00478	-9.85	<.0001	1.02166
ShelBad	1	-1.77565	0.17968	-9.88	<.0001	1.13394
ShelGood	1	3.06674	0.19074	16.08	<.0001	1.12169

Table 4: Variation inflation factor

Indeed, we could treat the multicollinearity aspect by using the ridge regression method (FigA3), but the result will mainly be the same. The analysis of the outputs below shows that at $\lambda = 0.4$ already, we have a first coefficient for shrinkage. That lead to retain the same variable as stated above: ***CompPrice, Income, Advertising, Price, Age, Shel_Bad*** and ***Shel_Good.*** Surely the $\hat{\beta}$ are biased but have the minimum mean square error.



Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Inflation de variance
Intercept	1	6.87726	0.86008	8.00	<.0001	0
CompPrice	1	0.09305	0.00647	14.37	<.0001	1.85098
Income	1	0.01901	0.00281	6.76	<.0001	1.06335
Advertising	1	0.12415	0.01716	7.23	<.0001	2.20872
Population	1	0.00013779	0.00053722	0.26	0.7978	1.15999
Price	1	-0.09066	0.00397	-22.82	<.0001	1.77503
Age	1	-0.04635	0.00485	-9.55	<.0001	1.04623
Education	1	-0.02007	0.02752	-0.73	0.4667	1.04305
ShelBad	1	-1.75173	0.18373	-9.53	<.0001	1.17707
ShelGood	1	3.10567	0.19316	16.08	<.0001	1.14197
Urban_Yes	1	0.08388	0.16503	0.51	0.6118	1.05689
US_Yes	1	-0.29850	0.22773	-1.31	0.1915	2.17339

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	11	1418.57233	128.96112	113.48	<.0001
Erreur	198	225.00857	1.13641		
Total sommes corrigées	209	1643.58091			

Root MSE	1.06602	R carré	0.8631
Moyenne dépendante	7.54267	R car. ajust.	0.8555
Coeff Var	14.13325		

Table 5: Ridge regression outputs

Now, a LASSO method could be applied to the data in the idea to seek for a better model selection.

Stime dei parametri

Parametro	DF	Stima
Intercept	1	7.246768
CompPrice	1	0.092113
Income	1	0.016885
Advertising	1	0.123094
Population	1	0.000254
Price	1	-0.092408
Age	1	-0.045313
Education	1	-0.027625
ShelveLoc_Bad	1	-1.920795
ShelveLoc_Good	1	2.957017
Urban_No	1	-0.128830
US_No	1	0.177970

Analisi della varianza

Origine	DF	Somma dei quadrati	Media quadratica	Valore F
Modello	11	2154.69686	195.88153	190.48
Errore	288	296.17099	1.02837	
Totale corretto	299	2450.86785		

Radice MSE	1.01409
Media dip.	7.49687
R-quadro	0.8792
R-quadro corr	0.8745
AIC	322.14635
AICC	323.41907
SBC	64.59174
CVEX PRESS	1.10293

MSE = 195.881

R_carré = 0.8792

R_carré_ajusté = 0.8745

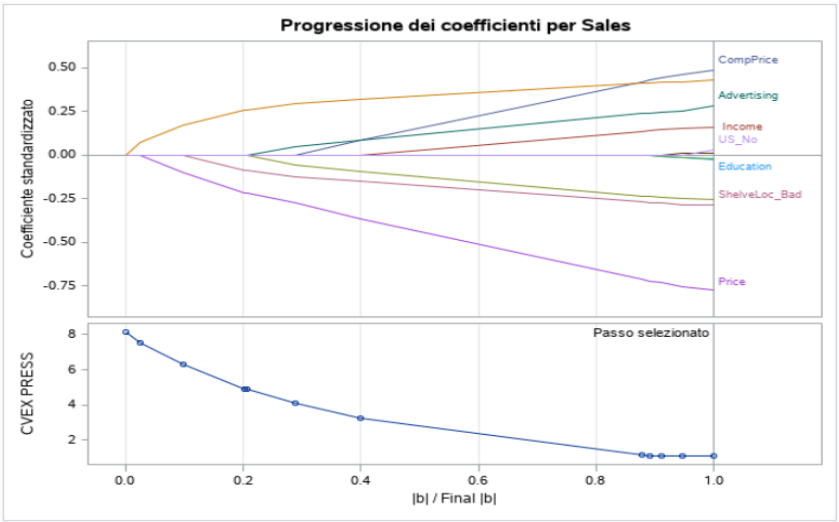


Table 6: Lasso regression output

The result of the LASSO method shows also that for a $\lambda = \mathbf{0.8}$, we have a relative minimum CVEX PRESS and the variables ***CompPrice, Income, Advertising, Price, Age, Shel_Bad*** and ***Shel_Good*** are the one that explain the sales mostly.

<u>Possible interaction between qualitative and quantitative variables</u>: the analysis of possible interaction between variables gives no interesting output. And we could suggest that there is no significative interactions between our selected variables.

<u>Model diagnostics:</u> :

Based on the variables selection done previously, our final model is :

$$\boldsymbol{sales} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} * \boldsymbol{CompPrice} + \boldsymbol{\beta_2} * \boldsymbol{Income} + \boldsymbol{\beta_3} * \boldsymbol{Advertising} + \boldsymbol{\beta_4} * \boldsymbol{Price} + \boldsymbol{\beta_5} * \boldsymbol{Age} + \boldsymbol{\beta_6} * \boldsymbol{ShelveLoc_{Bad}} + \boldsymbol{\beta_7} * \boldsymbol{ShelveLoc_{Good}} + \boldsymbol{\epsilon}$$

- <u>Linearity:</u> The analysis of the scatterplot of the residual *e* against the regressors (Fig2) confirmed that the model is linear.

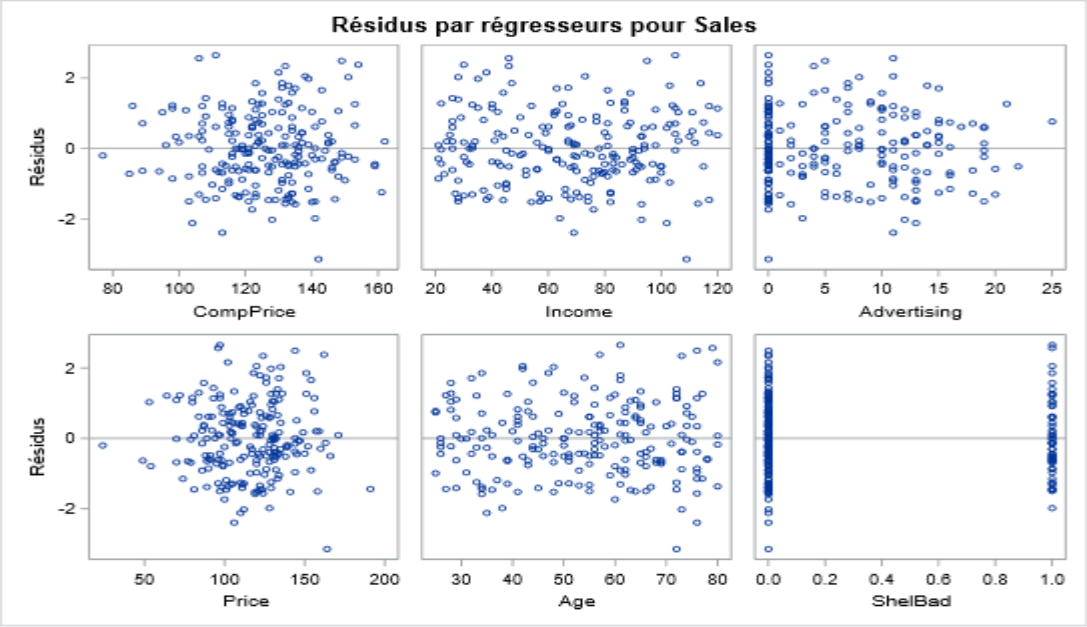


Fig2: scatterplot of the residual against the regressors

- <u>Normality of the residuals:</u> with the output of the fig3 below, we can see graphically that the residuals of the model are normally distributed such that $E(\epsilon) = 0$, **Shapiro-Wilk** test confirm that (table 7 below).

Test di posizione: Mu0=0				
Test	Statistica		P-value	
T di Student	t	0	Pr > |t|	1.0000
Segno	M	0	Pr >= |M|	1.0000
Rango con segno	S	24.5	Pr >= |S|	0.9221

Test di normalità				
Test	Statistica		P-value	
Shapiro-Wilk	W	0.995463	Pr < W	0.9908
Kolmogorov-Smirnov	D	0.0497	Pr > D	>0.1500
Cramer-von Mises	W-Qu	0.025536	Pr > W-Qu	>0.2500
Anderson-Darling	A-Qu	0.160865	Pr > A-Qu	>0.2500

Table 7: normality test

- Also, graphically the variance of errors terms tends to be constant showing that the hypotheses of homoscedasticity is respected. This is confirmed by the white test (table 6 below).

Test d'hétéroscédasticité					
Equation	Test	Statistique	DDL	Pr > khi-2	Variables
Sales	Test de White	37.48	32	0.2322	Croix de toutes les var

Table 6: homoskedasticity test of the error term

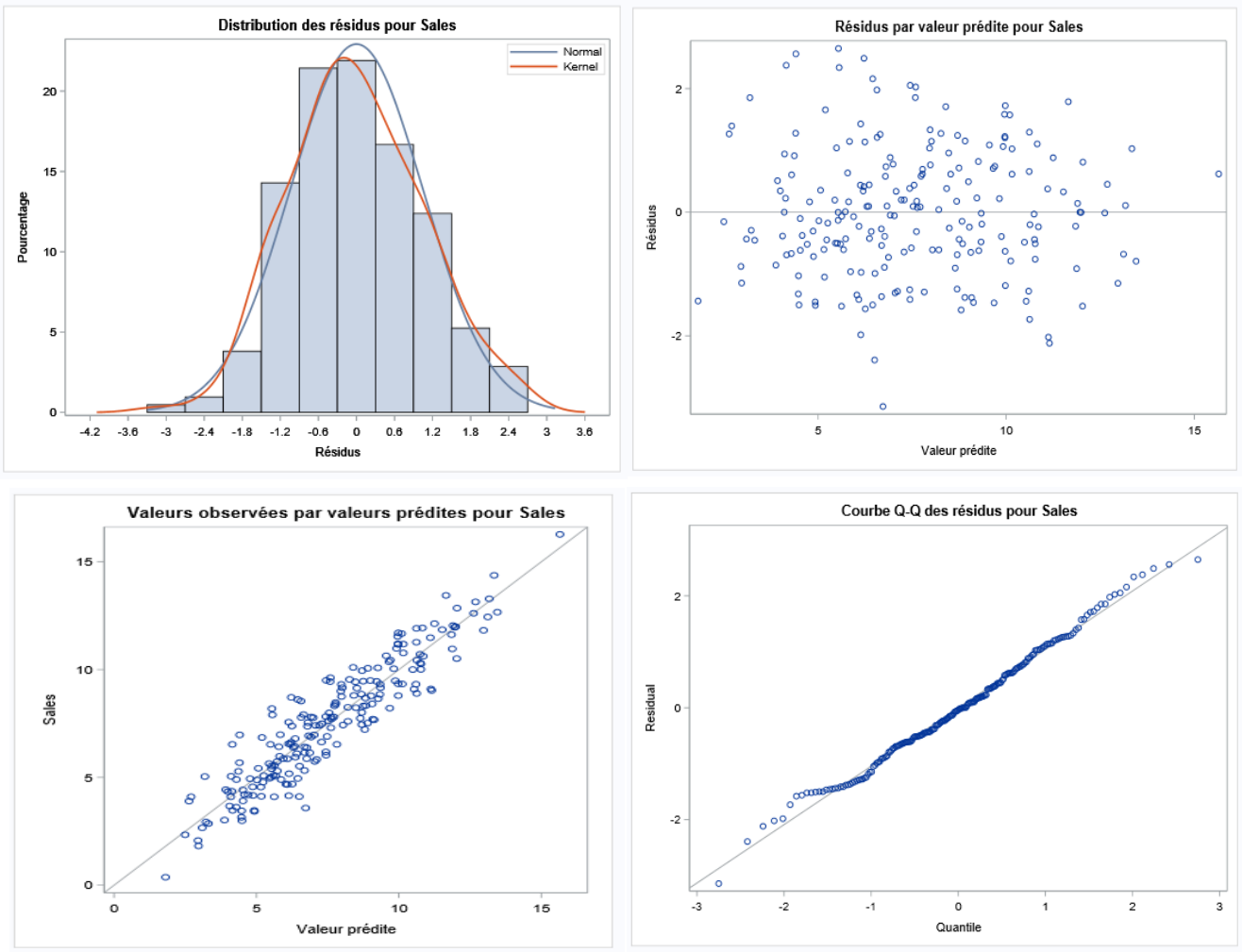


Fig 3: output of the residuals analysis

- <u>Outliers and influential observations:</u> The analysis of the outliers and influential observations is done throughout the plots below: we can see that nine observations are likely to be considered as outliers in the dataset.

We will start by looking out the outliers with respect to the explanatory variables by calculating the leverage h_{ii}. In the case in study, an observation will be considered as outlier if its leverage is larger than the threshold $2 * \frac{8}{210} = \mathbf{0,076}$. In our case, only observations 170, 190 and 208 have leverage greater than 0.076 and so are outliers with respect to the explanatory variables.

Obs.	lev
170	0.082950
190	0.080481
208	0.085089

Now, knowing that these observations are outliers, are they influential based on the DFFITS criterion?

Obs.	df
13	0.48221
15	0.41053
104	0.51124
138	0.55773
160	-0.49416
161	0.47094
163	0.48345
190	-0.41845
201	-0.87923

Our DFFITS criterion state that an observation is influential if $|\mathrm{DFFITS}_i| > \delta$ where $\delta = 2\sqrt{8/210} = \mathbf{0.39}$.

According to that criterion, the observations 13, 15, 104, 138, 160, 161, 163, 190, 201 are influential. Until now, only observations 190 is outlier and influential.

What is cook distance criterion said?

According to Cook's distance criterion under which an observation ii is considered influential both on the coefficients and on the predicted values there are no influential observations.

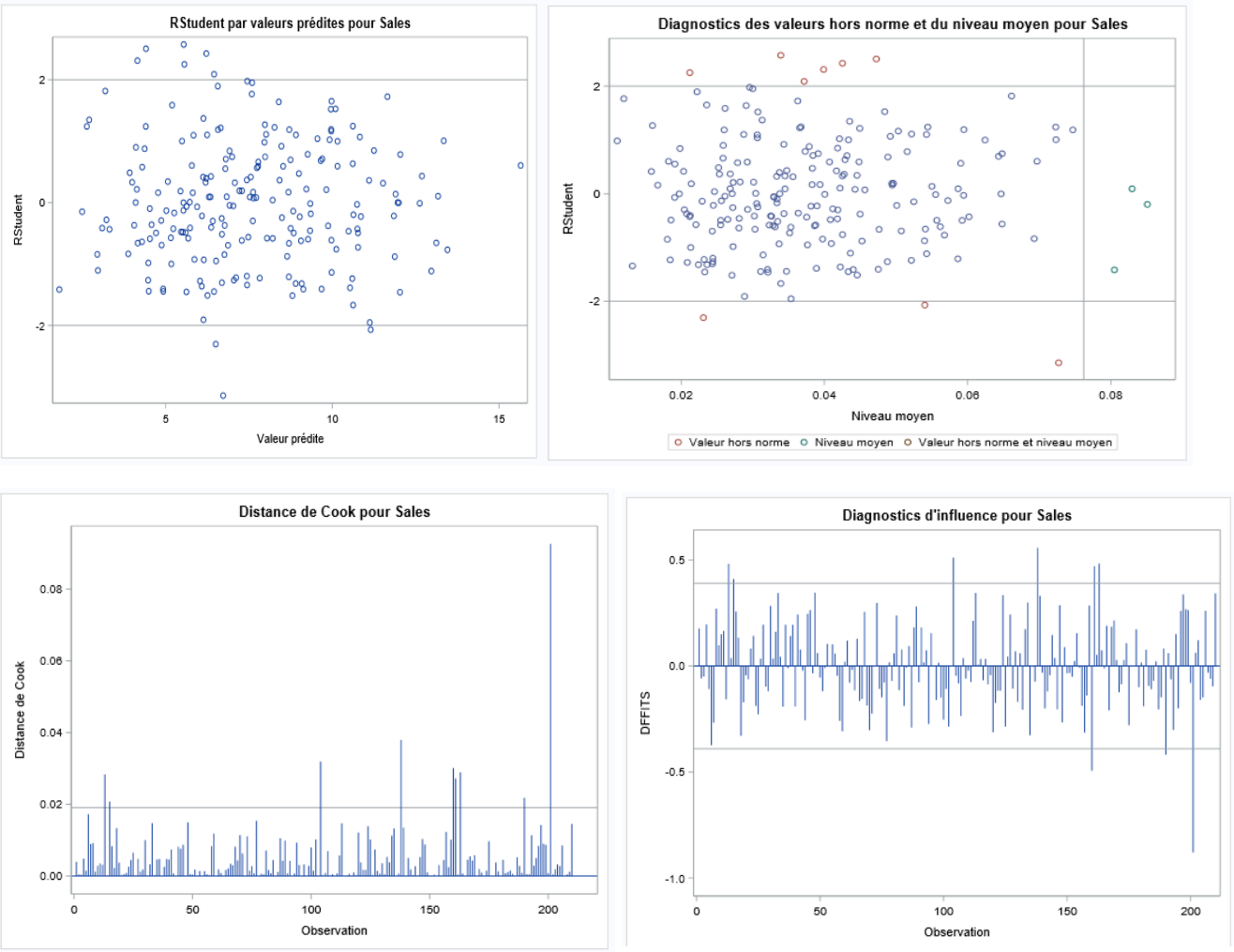


Fig 4: cook distance and dffits outputs

What are the influences on the regression coefficients?

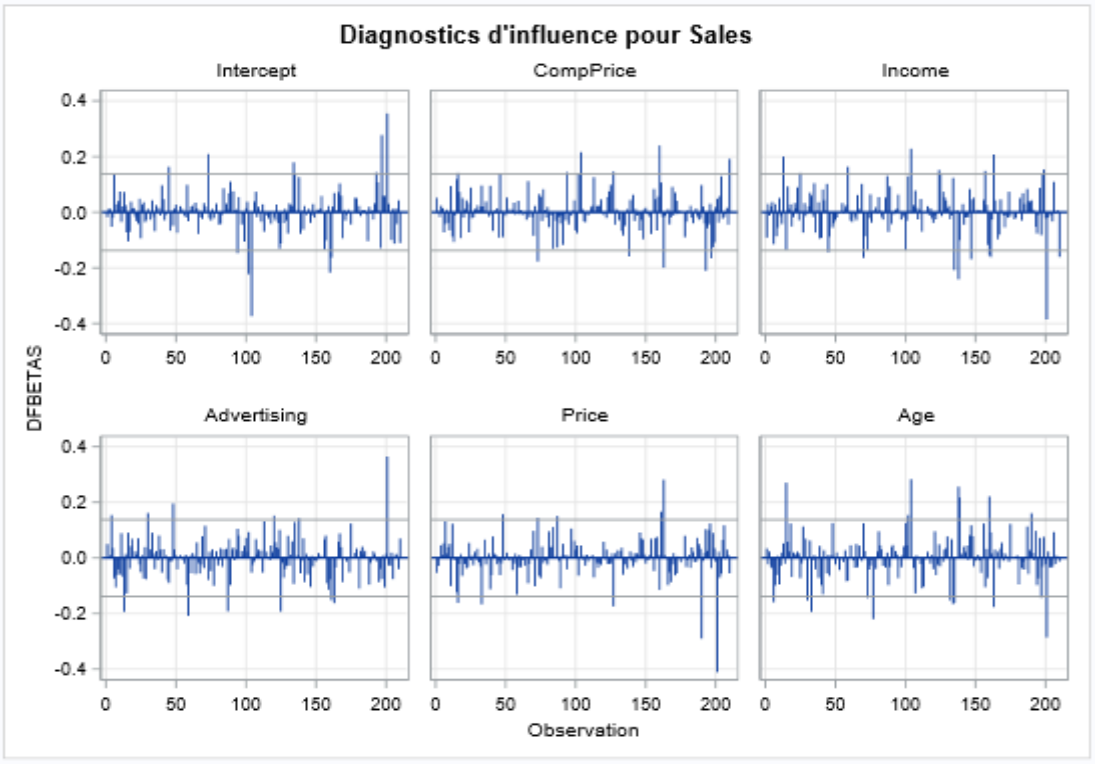


Fig 4: DFBETAS

The DFBETAS analysis shown the 200th observations could influential.

<u>Significance of the estimated coefficients and interpretations</u>: recall the table below where all the $\hat{\beta}$ are significant at 5% level.

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	6.71616	0.73742	9.11	<.0001
CompPrice	1	0.09265	0.00631	14.69	<.0001
Income	1	0.01906	0.00279	6.83	<.0001
Advertising	1	0.11008	0.01176	9.36	<.0001
Price	1	-0.09088	0.00391	-23.22	<.0001
Age	1	-0.04706	0.00478	-9.85	<.0001
ShelBad	1	-1.77565	0.17968	-9.88	<.0001
ShelGood	1	3.06674	0.19074	16.08	<.0001

Table 6: Estimated coefficients

First, $E(sales\,|\boldsymbol{X}=\mathbf{0}) = \beta_0 = 6.716$ showing that based on the model we specify, the expected sales is about **6716$** if all explanatory variables equal 0.

$\beta_1 = 0.093$ shows that when the price charged by the competitor in the region increases for **1 point**, the sales of the targeted store increase by **0.0093 point** and so obviously the increase of the price charged by the company resume by the decrease of the sales ($\beta_4 < 0$). Also, for **one unit** inverted in advertisement, the sales increases by **0.11 point**

In the case of qualitative variable, $\boldsymbol{\beta_6 = -1.775}$ and $\boldsymbol{\beta_7 = 3.067}$ shows that if the car seats shelve location in the store were judged as bad, the sales decreases by **1.775 point**. In contrast, when it's well located the sales increase by **3.067 point**.

Test for significance of some $\widehat{\beta}$:

- First, let's test the hypothesis that $\boldsymbol{\beta_1 + \beta_4 = 0}$ (Price and compPrice are opposite). Based on the result of the test in the table besides, the p-value = 0.71 so we don't reject the hypothesis that the price charged by the company and the price charged by the competitor are opposite.

Résultats du test 1 pour la variable dépendante Sales				
Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	1	0.15130	0.13	0.7146
Dénominateur	202	1.12809		

Table 7: test $\beta_1 + \beta_4 = 0$

- Now, we test $\boldsymbol{\beta_6 = \beta_7 = 0}$. (Table 8) As the p-value is less than 0.05, we reject the null hypothesis and consider that at least one of these variables (ShelveLoc_Good or ShelveLoc_Bad) is significant in our model.

Résultats du test 3 pour la variable dépendante Sales				
Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	2	279.73919	247.98	<.0001
Dénominateur	202	1.12809		

Table 8: test $\boldsymbol{\beta_6 = \beta_7 = 0}$.

Prediction interval:

For the prediction, we will use the test set, and compute the fitted value based on the model specified, and the prediction intervals for the predicted values. Now we generate a binary variable (well) whose 1 indicate that the observed sales value is in the predicted intervals, 0 otherwise. The frequencies of variable well in in the table below.

well	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
0	2	2.22	2	2.22
1	88	97.78	90	100.00

Obs.	Sales	predicted	lower_pred	upper_pred
26	5.55	7.68834	5.90492	9.47175
35	7.81	5.71990	3.91430	7.52550

Based on the output, only two observations (obs 26 and 35) are not in the predicted intervals. With 97.78% of observed value in the predicted intervals, we can say that the predicted power of our model is high.

Conclusion:

Sum up, the model

$$\boldsymbol{sales = \beta_0 + \beta_1 * CompPrice + \beta_2 * Income + \beta_3 * Advertising + \beta_4 * Price + \beta_5 * Age + \beta_6 * ShelveLoc_{Bad} + \beta_7 * ShelveLoc_{Good} + \epsilon}$$

Is well suitable to explain how the car seats sales evolve in the targeted region. To increase the sales of advertisement, the price charged comparairing the one charged by competitor and the shelve location of the produit in store have a significant part.

APPENDIX

Elenco alfabetico di variabili e attributi					
N.	Variabile	Tipo	Lungh	Formato	Formato di input
4	Advertising	Num	8	BEST12.	BEST32.
8	Age	Num	8	BEST12.	BEST32.
2	CompPrice	Num	8	BEST12.	BEST32.
9	Education	Num	8	BEST12.	BEST32.
3	Income	Num	8	BEST12.	BEST32.
5	Population	Num	8	BEST12.	BEST32.
6	Price	Num	8	BEST12.	BEST32.
1	Sales	Num	8	BEST12.	BEST32.
7	ShelveLoc	Alfanum	6	$6.	$6.
11	US	Alfanum	3	$3.	$3.
10	Urban	Alfanum	3	$3.	$3.

Fig A1: Variables of the dataset

Nombre dans le modèle	C(p)	R carré	Variables du modèle
7	6.5214	0.8614	CompPrice Income Advertising Price Age ShelBad ShelGood
8	6.8911	0.8625	CompPrice Income Advertising Price Age ShelBad ShelGood US_Yes
8	8.1815	0.8616	CompPrice Income Advertising Price Age Education ShelBad ShelGood
8	8.2081	0.8616	CompPrice Income Advertising Population Price Age ShelBad ShelGood
9	8.3142	0.8629	CompPrice Income Advertising Price Age Education ShelBad ShelGood US_Yes
8	8.3455	0.8615	CompPrice Income Advertising Price Age ShelBad ShelGood Urban_Yes
9	8.6446	0.8627	CompPrice Income Advertising Price Age ShelBad ShelGood Urban_Yes US_Yes
9	8.7916	0.8626	CompPrice Income Advertising Population Price Age ShelBad ShelGood US_Yes
9	9.9148	0.8618	CompPrice Income Advertising Population Price Age Education ShelBad ShelGood
9	10.0068	0.8617	CompPrice Income Advertising Population Price Age ShelBad ShelGood Urban_Yes
9	10.0081	0.8617	CompPrice Income Advertising Price Age Education ShelBad ShelGood Urban_Yes
10	10.0658	0.8631	CompPrice Income Advertising Price Age Education ShelBad ShelGood Urban_Yes US_Yes
10	10.2584	0.8629	CompPrice Income Advertising Population Price Age Education ShelBad ShelGood US_Yes
10	10.5319	0.8627	CompPrice Income Advertising Population Price Age ShelBad ShelGood Urban_Yes US_Yes
10	11.7180	0.8619	CompPrice Income Advertising Population Price Age Education ShelBad ShelGood Urban_Yes
11	12.0000	0.8631	CompPrice Income Advertising Population Price Age Education ShelBad ShelGood Urban_Yes US_Yes
6	50.8880	0.8293	CompPrice Advertising Price Age ShelBad ShelGood
7	51.8335	0.8300	CompPrice Advertising Price Age ShelBad ShelGood US_Yes
7	51.9530	0.8299	CompPrice Advertising Price Age ShelBad ShelGood Urban_Yes
7	52.2182	0.8298	CompPrice Advertising Price Age Education ShelBad ShelGood
7	52.5897	0.8295	CompPrice Advertising Population Price Age ShelBad ShelGood
8	52.7665	0.8308	CompPrice Advertising Price Age ShelBad ShelGood Urban_Yes US_Yes
8	52.9028	0.8307	CompPrice Advertising Price Age Education ShelBad ShelGood US_Yes
8	53.2961	0.8304	CompPrice Advertising Price Age Education ShelBad ShelGood Urban_Yes
8	53.5974	0.8302	CompPrice Advertising Population Price Age ShelBad ShelGood Urban_Yes

FigA2: Mallow output.

Programme SAS:

```
FILENAME myfile '/home/u48508240/PROJET_LSTAT2120_2021/carseats.csv';
/*importer les données*/
PROC IMPORT DATAFILE=myfile
        DBMS=CSV
        OUT=WORK.CAR
        replace;
        delimiter=';' ;
        GETNAMES=YES;
RUN;
proc print data= WORK.CAR;
run;
/*histogramme des variables*/

ods graphics on;
proc univariate data=CAR;
  var Education CompPrice Income Population Price;
  histogram;
run;
/* statistiques descriptives*/
proc means data=CAR N mean std var skewness kurtosis maxdec=2 ;
run;
/* on remarque une corrélation assez importante entre Price et CompPrice (0.58),
et une corrélation (non-négligeable) négative entre Sales/Price(-0.44) et positive entre
 Sales/Advertising(0.269) */

proc corr data=CAR;
run;
ods graphics off;
```

```
data CAR1;
set CAR;
if ShelveLoc='Bad' then Shel_Bad=1; else Shel_Bad=0;
if ShelveLoc='Good' then Shel_Good=1; else Shel_Good=0;
if ShelveLoc='Medium' then Shel_Med=1; else Shel_Med=0;

if Urban='Yes' then Urb_Yes=1; else Urb_Yes=0;
if Urban='No' then Urb_No =1; else Urb_No =0;

if US='Yes' then US_Yes =1; else US_Yes =0;
if US='No' then US_No =1; else US_No =0;

run;



/* CAR2= sous-ensemble de 300 observations pour le modèle*/
/* CAR_PRED = sous-ensemble de 100 observations reservées pour la prédiction*/

data temp2;
set CAR1;
n=ranuni(8);
run;
proc sort data=temp2;

  by n;
run;

data CAR2 CAR_PRED;

set temp2 nobs=nobs;
if _n_<=.75*nobs then output CAR2;
else output CAR_PRED;
run;
```

```
/* création des sous-ensembles test(30%) et train(70%)*/
data temp;
set CAR2;
n=ranuni(8);
run;

proc sort data=temp;

 by n;
run;

data train test;

set temp nobs=nobs;
if _n_<=.7*nobs then output train;
else output test;
run;

ods graphics on;

/* stepwise*/
proc reg data=train outest=sss tableout plots=none;
model Sales= CompPrice Income Advertising Population Price Age Education Shel_Bad  Shel_Good Urb_Yes US_Yes / selection=stepwise vif collinoint spec ;
run; quit;



proc print data=sss; run; quit;

/* ridge regression*/

proc reg data=train ridge= 0 to 1 by 0.1 outest=ridgest;
model Sales= CompPrice Income Advertising Population Price Age Education Shel_Bad  Shel_Good Urb_Yes US_Yes / vif spec dw;
```

```
run;

/*Lasso*/
proc glmselect data=CAR2 plots(stepaxis=normb)=all seed=123 outdesign=lassgest;
class ShelveLoc Urban US;
model Sales= CompPrice Income Advertising Population Price Age Education ShelveLoc Urban US /
selection=lasso(stop=none choose=cvex) cvmethod=random(5)  ;

run;

ods graphics off;

/* Test linear combination Comprice & Price, and some variables*/

proc reg data=train plots=NONE;
model Sales= CompPrice Income Advertising Population Price Age Education Shel_Bad  Shel_Good
Urb_Yes US_Yes;
test CompPrice + Price =0;
test Income = 0, Price = 0, Shel_Good = 0;

run;

/*prediction interval*/

proc reg data=test plots=none;
  model Sales= CompPrice Income Advertising Price Age Shel_Bad  Shel_Good/ vif spec stb clb  ;
  output out=stdres p=predicted r=resid ucl=upper_pred lcl=lower_pred;
run;

/*test de normalité des résidus (Shapiro-Wilk)*/
proc univariate data=stdres normal;
var resid;
run;
```

```
proc print data=test;
var Sales predicted  lower_pred upper_pred;
run;


/*test d'homoscdasticité*/
proc reg data=train;
model Sales= CompPrice Income Advertising Price Age Shel_Bad  Shel_Good / spec;
run;

ods graphics on;
proc model data=train;
parms b0 b1 b2 b3 b4 b5 b6 b7;
Sales= b0 + b1*CompPrice + b2*Income + b3*Advertising + b4*Price + b5*Age + b6*ShelBad +
b7*ShelGood;
fit sales/white;
run;

ods graphics off;


/*outliers*/

/*DFFITS*/

/* |DFFITS| = 2*sqrt(8/210) = 0.41*/

proc reg data=train plots(only)=dffits;
model Sales= CompPrice Income Advertising Price Age Shel_Bad  Shel_Good /spec;
/*plot h.*obs.; */
output out=outliers dffits=df;
run;
```

```
proc print data=outliers;
var lev;
where lev >0.076;
run;

proc print data=outliers;
 var df;
 where df > 0.34 or df < -0.34;
run;
```