Projet d'économétrie

Les facteurs de ventes de Walmart

Joudy BENKADDOUR et Ryhan CHEBREK

Partie 1

« Je déclare su l'honneur que ce mémoire a été rédigée de ma main, sans aide extérieure non autorisée, qu'il n'a pas été présenté auparavant pour évaluation et qu'il n'a jamais été publié, dans sa totalité ou en partie.
Toutes parties, groupes de mots ou idées, aussi limités soient-ils, y compris des tableaux, graphiques, cartes etc. qui sont empruntés ou qui font référence à d'autres sources bibliographiques sont présentés comme tels, sans exception aucune. »

Préambule

Dans toute la suite du projet, le ***seuil de risque de rejet a été fixé à 5%*** *ce qui semble être une hypothèse convenable et conforme à nos données économiques.*

Le type de modèle étudié pour ce projet est le ***modèle en série temporelle*** *pour des variables allant du 5 février 2010 au 26 octobre 2012.*

L'intégralité des données ont été récupérées sur le site de base de données Kaggle à l'adresse https://www.kaggle.com/yasserh/walmart-dataset. Le choix du sujet a fait l'objet d'un arbitrage entre une pluralité de bases de données étudiées sur leur pertinence. Dans notre cas, c'est le nombre de ventes quotidiennes comme variable qui a été retenu pour être expliqué par rapport au type de semaine (vacances ou non), la température, le prix du gasoil (dans la région) ainsi que deux variables macroéconomique telles que le taux d'inflation et de chômage tous prélevés sur la semaine de référence.

Seul le magasin numéro 1 dans la base de données a été utilisé pour ce modèle. On compte n=143 semaines de référence, pour k=5 (+ une constante) variables expliquant le modèle.

Ces variables explicatives semblent à première vue être assez reliées au phénomène qu'elles cherchent à expliquer à savoir les ventes. L'objet de ce projet sera donc d'évaluer et de mesurer à quel point ce modèle semble être pertinent et explicatif selon plusieurs tests et processus significatifs.

I. Estimation du modèle initial

1) Test de significativité globale

Modèle 1: MCO, utilisant les observations 1-143
Variable dépendante: Ventes_quotidiennes_y

	coefficient	éc. type	t de Student	p. critique	
const	-2,42786e+06	1,75296e+06	-1,385	0,1683	
Type_de_semaine_~	89375,5	49338,2	1,811	0,0723	*
Temperature_a2	-2160,42	922,044	-2,343	0,0206	**
Prix_du_gasoil_a3	-24337,0	47335,4	-0,5141	0,6080	
Taux_inflation_a4	16631,6	6786,09	2,451	0,0155	**
Taux_de_chomage_~	80209,4	58726,5	1,366	0,1742	

Moyenne var. dép.	1555264	Éc. type var. dép.	155980,8
Somme carrés résidus	2,94e+12	Éc. type régression	146453,4
R2	0,149471	R2 ajusté	0,118430
F(5, 137)	4,815260	P. critique (F)	0,000436
Log de vraisemblance	-1900,752	Critère d'Akaike	3813,503
Critère de Schwarz	3831,280	Hannan-Quinn	3820,727
rho	0,136457	Durbin-Watson	1,723788

$H_0 : \hat{a}_0=\hat{a}_1=\ldots=\hat{a}_i=\ldots= 0$
H_1 : Il existe au moins un coefficient significativement différent de 0 : $\hat{a}_i \neq 0$

Ce test permet de montrer qu'il existe des variables exogènes expliquant les ventes du Walmart.

Pour effectuer le test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider

Dans notre modèle, on obtient F* = 4.815 et la p-value est de 0.000436.

Ainsi on rejette l'hypothèse H_0 au seuil de 5% donc il existe au moins un coefficient significativement différent de 0.

Notre modèle est donc utilisable car au moins une variable peut expliquer le volume de vente du magasin.

2) Test de significativité individuelle pour toutes les variables

$H_0 : \hat{a}_i = 0$ $\hat{a}_i$ significativement = 0 : La variable n'explique pas les ventes
$H_1 : \hat{a}_i \neq 0$ $\hat{a}_i$ significativement différent de 0 : La variable explique les ventes

Ce test permet de montrer pour chaque variable si elle explique les ventes. Nous allons le faire pour les deux premières variables c'est-à-dire le type de semaine (1 si vacances et 0 sinon) et la température.

Pour effectuer le test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider

Pour notre modèle :

Pour $\hat{a}_1$: La probabilité critique est de 7,23% donc on ne rejette pas H_0 au seuil de 5%
Pour $\hat{a}_2$: La probabilité critique est de 2,03% donc on rejette H_0 au seuil de 5%

Ainsi on conclut que le type de semaine (vacances ou non) explique significativement le volume de vente tandis que la température n'explique pas significativement le modèle.

3) <u>Test de significativité à valeur fixée (unilatéral)</u>

$H_0 : a_4 = 10000$
$H_1 : a_4 > 10000$

Ce test nous permet de savoir si l'on peut contraindre une variable à une valeur sans nuire à la qualité du modèle.

Pour effectuer ce test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Restrictions linéaires => b[5]=10000 => Valider

Pour notre modèle : La probabilité critique vaut 33,02%.

Or Gretl calcule la p-value en bilatéral donc on divise par 2 pour obtenir la valeur de la p-value en unilatéral.

p-value en bilatéral = 33,02% donc la p-value en unilatéral = 16,01%, ainsi on accepte H_0 au seuil de 5%.

On en conclu que forcer le coefficient $\hat{a}_4 = 10000$ ne diminue pas drastiquement la qualité du modèle.

```
Restriction:
 b[Taux_inflation_a4] = 10000

Statistique de test: F(1, 137) = 0,954984, avec p. critique = 0,330176

Estimations contraintes:

                      coefficient    éc. type     t de Student  p. critique
  ------------------------------------------------------------------------
  const                -750954      358145         -2,097        0,0378   **
  Type_de_semaine_a1     94241,0     49078,3         1,920        0,0569   *
  Temperature_a2         -2338,68      903,671      -2,588        0,0107   **
  Prix_du_gasoil_a3       7869,81    33971,9         0,2317       0,8171
  Taux_inflation_a4      10000,0         0,000000    NA           NA
  Taux_de_chomage_a5     36012,3     37455,0         0,9615       0,3380

  Écart-type de la régression = 146429
```

4) Test de combinaison linéaire (unilatéral)

$H_0 : \hat{a}_5 - 5*\hat{a}_4 = 0$
$H_1 : \hat{a}_5 - 5*\hat{a}_4 < 0$

Ce test nous permet de vérifier si l'on peut trouver des relations linéaires entre les variables sans nuire à la qualité du modèle.

Pour effectuer ce test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Restrictions linéaires => b[6]-5*b[5]=0 => Valider

```
Restriction:
 b[Taux_de_chomage_a5] - 5*b[Taux_inflation_a4] = 0

Statistique de test: F(1, 137) = 0,00567907, avec p. critique = 0,940039

Estimations contraintes:

                       coefficient        éc. type      t de Student  p. critiq
  ------------------------------------------------------------------------------
  const                  -2,48995e+06      1,54168e+06    -1,615        0,1086
  Type_de_semaine_a1     89154,1          49072,9           1,817        0,0714
  Temperature_a2         -2151,98           911,914        -2,360        0,0197
  Prix_du_gasoil_a3     -23846,4          46716,4          -0,5105       0,6106
  Taux_inflation_a4      16779,2           6473,68          2,592        0,0106
  Taux_de_chomage_a5     83896,2          32368,4           2,592        0,0106
```

Ici la p-value vaut 94%.
Or Gretl calcule la p-value bilatéral donc on divise par 2 pour obtenir la valeur de la p-value unilatéral.
La p-value en bilatéral = 94% donc la p-value unilatéral = 47%, ainsi on accepte H0 au seuil de 5%.

On en conclu que forcer le coefficient â5 = 5*â4 ne diminue pas drastiquement la qualité du modèle.

II. Test par analyse de la variance

1) Test d'égalité conjointe de 2 coefficients

H_0 : $\hat{a}_1 = 90000$ et $\hat{a}_2 = -2000$
H_1 : 3 possibilités

Il existe 3 causes de rejets possibles de H_0 : $\hat{a}_1 \neq 90000$ ou $\hat{a}_2 \neq -2000$ ou $\hat{a}_1 \neq 90000$ et $\hat{a}_2 \neq -2000$.

Ce test nous permet de savoir si l'on va pouvoir conjointement fixer deux variables sans nuire à la qualité du modèle.

Pour effectuer ce test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Restrictions linéaires => b[2] = 90000 et b[3] = -2000 => Valider

```
Ensemble de restrictions
 1: b[Temperature_a2] = -2000
 2: b[Type_de_semaine_a1] = 90000

Statistique de test: F(2, 137) = 0,015251, avec p. critique = 0,984866

Estimations contraintes:

                      coefficient        éc. type      t de Student
  ------------------------------------------------------------------
  const                 -2,50156e+06     1,67802e+06    -1,491
  Type_de_semaine_a1    90000,0          0,000000       NA
  Temperature_a2        -2000,00         0,000000       NA
  Prix_du_gasoil_a3     -26455,5         45165,8        -0,5857
  Taux_inflation_a4     16879,1          6542,54         2,580
  Taux_de_chomage_a5    82321,1          56676,5         1,452

  Écart-type de la régression = 145412
```

Ainsi on a une p-value = 98,48%, alors on accepte H0 au seuil de 5%.

Cela signifie que l'on peut garder â1 = 90000 et â2 = -2000 sans véritablement dégrader le modèle.

2) Test d'ajout de variable

$H_0 : SCR = SCR_1$
$H_1 : SCR \neq SCR_1$

Ce test nous permet de savoir si l'on peut ajouter ou supprimer des variables sans nuire à la qualité du modèle.

On a décidé d'omettre les variables : inflation, température, type de semaine qui sont les coefficients avec les plus grandes probabilités critiques donc les variables qui expliquent le mieux les ventes.

Pour effectuer ce test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Omettre les variables => Sélectionner inflation, température et type de semaine => Valider

```
Test sur le Model 1 :

  Hypothèse nulle : les paramètres de la régression sont nuls pour les variables
    Type_de_semaine_a1, Temperature_a2, Taux_inflation_a4
  Statistique de test: F(3, 137) = 7,10801, p. critique 0,000178886
  L'omission de variables améliore 0 parmi 3 critères d'information.

Modèle 2: MCO, utilisant les observations 1-143
Variable dépendante: Ventes_quotidiennes_y

                        coefficient     éc. type  t de Student  p. critique
  ------------------------------------------------------------------------
  const                   1,57928e+06  374566       4,216       4,43e-05   ***
  Prix_du_gasoil_a3     36831,0          35655,0     1,033       0,3034
  Taux_de_chomage_~    -18737,6          39702,7    -0,4719      0,6377

Moyenne var. dép.      1555264   Éc. type var. dép.    155980,8
Somme carrés résidus   3,40e+12  Éc. type régression   155743,1
R2                     0,017087  R2 ajusté             0,003045
F(2, 140)              1,216871  P. critique (F)       0,299268
Log de vraisemblance  -1911,095  Critère d'Akaike      3828,190
Critère de Schwarz     3837,078  Hannan-Quinn          3831,802
rho                    0,288740  Durbin-Watson         1,417124
```

On a donc une p-value très faible, on rejette l'hypothèse H_0 au seuil de 5%.
Assez logiquement, si l'on enlève les variables explicatives les plus pertinentes, le modèle détériore drastiquement la qualité du modèle.

3) Test de Chow

Ce test nous permet de savoir si le modèle reste stable lorsqu'on scinde l'échantillon.

"2 août 2011 : USA - Après des semaines de bataille, le Congrès autorise un relèvement du plafond de la dette (plus de 14.500 mds de dollars), évitant au pays un défaut de paiement. Avec un taux de chômage de 9,2% et une croissance faible (1,3%), les Etats-Unis se voient privés pour la première fois de leur "AAA" par l'agence de notation Standard and Poor's." (20 minutes).

Nous avons décidé de scinder notre modèle en 2 parties : la première avant le 2 août 2011 et la deuxième partie après cette date. Cette date nous semble pertinente car il nous permet de voir si les décisions du congrès ont eu des conséquences macroéconomiques sur les ventes et l'influence des coefficients des variables explicatives.

$H_0 : SCR - (SCR_1 + SCR_2) = 0$
$H_1 : SCR - (SCR_1 + SCR_2) \neq 0$
Ici n1 = 79 et n2 = 64

Pour effectuer ce test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Test de Chow => Scindé à 82 => Valider

```
Régression augmentée pour le test de Chow
MCO, utilisant les observations 1-143
Variable dépendante: Ventes_quotidiennes_y

                      coefficient         éc. type      t de Student  p. critique
  ------------------------------------------------------------------------------
  const                514060              5,70291e+06     0,09014      0,9283
  Type_de_semaine_~     88743,5           70562,3           1,258        0,2108
  Temperature_a2        -1799,00           1175,05         -1,531        0,1282
  Prix_du_gasoil_a3      7304,66         124757             0,05855      0,9534
  Taux_inflation_a4      6626,88          29670,0           0,2234       0,8236
  Taux_de_chomage_~    -38924,3          153385            -0,2538       0,8001
  splitdum                -2,03063e+07       9,55097e+06    -2,126        0,0354    **
  sd_Type_de_semai~      6371,69          99079,2           0,06431      0,9488
  sd_Temperature_a2       842,756          2243,78          0,3756       0,7078
  sd_Prix_du_gasoi~    -50790,8          161696            -0,3141       0,7539
  sd_Taux_inflatio~     77690,7          41986,1           1,850        0,0665    *
  sd_Taux_de_choma~    450034            218105            2,063        0,0411    **

Moyenne var. dép.      1555264   Éc. type var. dép.    155980,8
Somme carrés résidus   2,80e+12  Éc. type régression   146073,6
R2                     0,190933  R2 ajusté             0,122996
F(11, 131)             2,810443  P. critique (F)       0,002540
Log de vraisemblance -1897,178   Critère d'Akaike      3818,357
Critère de Schwarz     3853,911  Hannan-Quinn          3832,804
rho                    0,052365  Durbin-Watson         1,890970

Test de Chow pour rupture structurelle à l'observation 82
  F(6, 131) = 1,11887 avec p. critique 0,3548
```

On obtient une p-value de 35,48%, ainsi on accepte H_0 au seuil de 5%, cela signifie que les coefficients de chaque variable peuvent être considérés égaux avants et après la date du 11 août 2011. Le modèle est stable avant et après ce 11 août.

Ainsi, la décision du congrès n'a pas eu de conséquences directes sur les ventes du Walmart et sur l'importance des coefficients des variables explicatives.

4) <u>Test de contrainte sur les coefficients</u>

H_0 : $\hat{a}_1=\hat{a}_5$ et $\hat{a}_4=15000$
H_1 : Plusieurs possibilités

Ce test permet de savoir si l'on peut contraindre des coefficients sans nuire au modèle.

Il y a 3 raisons de rejeter H_0 : $\hat{a}_1$ significativement différent de $\hat{a}_5$ ou $\hat{a}_4$ significativement différent de 15000 ou bien les 2 à la fois.

Pour effectuer ce test : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Omettre les variables => Sélectionner inflation, température et type de semaine => Valider

```
Ensemble de restrictions
 1: b[Type_de_semaine_a1] - b[Taux_de_chomage_a5] = 0
 2: b[Taux_inflation_a4] = 15000

Statistique de test: F(2, 137) = 0,0862823, avec p. critique = 0,917385

Estimations contraintes:

                       coefficient       éc. type     t de Student  p. critique
  ------------------------------------------------------------------------------
  const                  -2,08262e+06  296013          -7,036       8,26e-011  ***
  Type_de_semaine_a1     77235,6          29123,2          2,652       0,0089     ***
  Temperature_a2         -2235,80           893,087       -2,503       0,0135     **
  Prix_du_gasoil_a3      -13212,7         32422,4         -0,4075      0,6843
  Taux_inflation_a4       15000,0             0,000000    NA           NA
  Taux_de_chomage_a5     77235,6          29123,2          2,652       0,0089     ***

  Écart-type de la régression = 145487
```

Ainsi, notre test nous renvoie une p-value = 91,73%, on accepte donc H_0 au seuil de 5%
Cela signifie que nos hypothèses ne détériorent pas significativement le modèle.

III. Intégration d'une variable indicatrice

Notre modèle possède une variable muette par défaut : le type de semaine de référence (vacances ou non) qui prend comme valeur 1 si c'est une période de vacances, 0 sinon. Précédemment, on a observé la probabilité critique associé à cette variable qui était à hauteur de 7,23% donc on ne rejette pas H_0 au seuil de 5%.

Ce qui implique que le coefficient associé à la variable muette (a1) est significativement proche de 0 et qu'ainsi, avoir une semaine de vacances comme référence n'impacte pas significativement le nombre de ventes.

IV. Elimination de variable

Nous allons éliminer les variables des moins significatives aux plus significatives donc des probabilités critiques les plus élevés au moins élevés.

Pour cela : Modèle => MCO => Définir variables dépendantes et explicatives => Valider => Tests => Omettre les variables => Estimations séquentielle 0.05 => Valider

```
Élimination séquentielle à l'aide de alpha bilatéral = 0,05

 Elimination de Prix_du_gasoil_a3 (p. critique 0,608)
 Elimination de Taux_de_chomage_a5 (p. critique 0,205)
 Elimination de Type_de_semaine_a1 (p. critique 0,051)

Test sur le Model 1 :

  Hypothèse nulle : les paramètres de la régression sont nuls pour les variables
    Type_de_semaine_a1, Prix_du_gasoil_a3, Taux_de_chomage_a5
  Statistique de test: F(3, 137) = 1,90985, p. critique 0,130878
  L'omission de variables améliore 3 parmi 3 critères d'information.

Modèle 2: MCO, utilisant les observations 1-143
Variable dépendante: Ventes_quotidiennes_y

                    coefficient   éc. type   t de Student  p. critique
  ------------------------------------------------------------------
  const              -233190      616327       -0,3784       0,7057
  Temperature_a2       -2768,86     876,980    -3,157        0,0020    ***
  Taux_inflation_a4     9155,62    2872,37      3,187        0,0018    ***

Moyenne var. dép.      1555264   Éc. type var. dép.    155980,8
Somme carrés résidus  3,06e+12   Éc. type régression   147874,2
R2                    0,113901   R2 ajusté             0,101242
F(2, 140)             8,997947   P. critique (F)       0,000211
Log de vraisemblance -1903,681   Critère d'Akaike      3813,362
Critère de Schwarz    3822,251   Hannan-Quinn          3816,974
rho                   0,189941   Durbin-Watson         1,615601
```

On obtient qu'il ne reste que 2 variables qui sont belle et bien significatives avec des p-value bien inférieur à 5%.

V. Conclusion

On peut dorénavant conclure avec tous les tests fait précédemment que le modèle obtenu à la fin de la précédente partie est le plus pertinent avec deux variables explicatives tout en minimisant l'erreur.

Le modèle est le suivant :

*Vente par semaine = -233190 - 2768,86 * Température + 9155,62 * Taux d'inflation*

On peut en conclure que les ventes sont fortement influées par la température et le taux d'inflation. Avec les coefficients, on peut donc avancer qu'une hausse des températures cause une baisse des ventes tandis qu'une hausse de l'inflation conduit à une hausse des ventes.

La relation entre température et vente du magasin est un peu biaisée car les périodes de fêtes où l'on consomme le plus sont en hiver.
Tandis que la relation entre inflation et volume de vente semble logique.
Nous pouvons aussi avancer qu'avec les probabilité obtenu le taux d'inflation est plus important que la température pour expliquer les ventes.

Le modèle est donc cohérent.