

## Chapitre 7, exercice 1

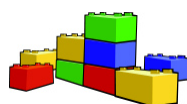
### Instructions pour employer CH07EX01.XLS

*Le fichier CH07EX01.XLS comporte un exercice de base destiné à tous les apprenants et un exercice avancé réservé aux seuls apprenants de la version avancée du cours.*

*Le répertoire CH07EX01 comporte un exercice de base destiné à tous les apprenants*

#### Exercice de base (Pour tous les utilisateurs du cours)

##### Préalable



Le chapitre 7 du cours de base doit avoir été suivi jusqu'à la page 46 pour la partie 1, jusqu'à la page 57 pour la partie 2, jusqu'à la page 59 pour la partie 3, jusqu'à la page 65 pour la partie 4, jusqu'à la page 71 pour la partie 5, jusqu'à la page 102 pour la partie 6, jusqu'à la page 192 pour la partie 7.

##### Objectif



Le but de l'exercice est d'introduire — ou de rappeler pour les apprenants qui ont déjà bénéficié d'un cours de statistique de base — les principaux éléments de la régression linéaire multiple par la méthode des moindres carrés.

##### Données



Les données sont relatives à une petite enquête pilote sur les prix des habitations en fonction des caractéristiques telles que la superficie habitable et la situation. Ces données sont en très petit nombre de manière à permettre de vérifier les calculs et d'appliquer aisément plusieurs approches.



### Structure de l'exercice

L'exercice comporte sept parties :

- Dans la partie 1, le but de l'exercice est d'introduire la régression linéaire multiple à partir des résultats fournis par l'outil Regression de la boîte à outils d'analyse de Microsoft Excel.
- Dans la partie 2, le but de l'exercice est d'introduire les mesures de qualité de l'ajustement que sont la somme des carrés des résidus et le coefficient de détermination  $R^2$ .
- Dans la partie 3, le but de l'exercice est d'expérimenter la régression linéaire multiple avec le logiciel Time Series Expert for Windows afin d'examiner différents jeux de variables explicatives.
- Dans la partie 4, le but de l'exercice est d'introduire les mesures de qualité de l'ajustement que sont le coefficient de détermination  $R^2$  corrigé ou  $\bar{R}^2$  ainsi que la variance résiduelle et l'écart-type résiduel.
- Dans la partie 5, les principaux aspects d'inférence statistique sont examinés : tests sur les coefficients, test global sur le modèle.
- Dans la partie 6, les résidus sont inspectés dans le but de prendre éventuellement le modèle en défaut par rapport au respect des conditions d'application.
- Dans la partie 7, un jeu de données légèrement étendu est examiné avec l'utilisation d'une variable de situation à 3 valeurs au lieu de 2, ce qui oblige à introduire deux variables binaires dans le modèle.



**Partie 1** Dans cette partie, le but de l'exercice est d'introduire la régression linéaire multiple à partir des résultats fournis par l'outil Regression de la boîte à outils d'analyse de Microsoft Excel.

### 1.1 PRÉSENTATION DES DONNÉES

Commençons par examiner les données qui sont relatives au prix de vente de 10 maisons. Nous allons essayer d'expliquer ce prix en fonction de deux caractéristiques : la superficie habitable et la situation. La variable de prix appelée PRICE est exprimée en milliers d'euros. La variable de superficie habitable appelée HAB est exprimée en m<sup>2</sup>. La variable de situation appelée SIT est codée comme suit : elle vaut 1 si l'habitation est dans un quartier résidentiel et 0 si l'habitation est dans le centre ville. Les données sont présentées dans un tableau.

⇒ Pour atteindre ce tableau, cliquez sur l'onglet de la feuille Main, pressez F5 et sélectionnez TABLE2. Vous pouvez aussi cliquer sur le lien prévu en haut de la feuille principale Main : «Pour les valeurs ajustées et résidus ».

Pour demander la régression multiple de PRICE en fonction de HAB et SIT, procédez comme suit. Notez que ces instructions ont déjà été données dans le chapitre 2, exercice 1, partie 3.

### 1.2 UTILISATION DE L'OUTIL DE RÉGRESSION

⇒ Cliquez sur le menu Tools ⇒ Data Analysis ⇒ Regression (descendez l'ascenseur à cet effet). Cliquez sur OK.

#### Remarque

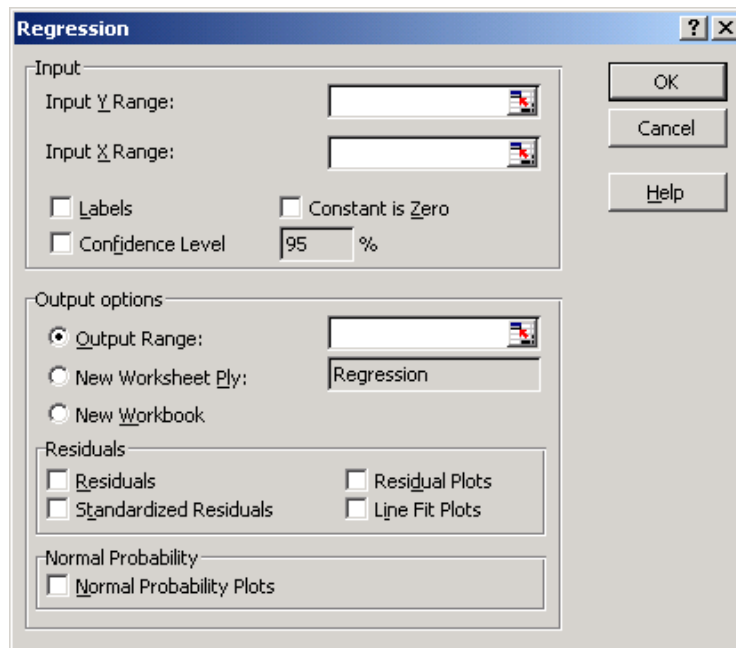


Il est fréquent que les outils d'analyse ne soient pas installés parce qu'ils ne le sont pas automatiquement au moyen du programme d'installation, à moins qu'une installation complète ait été demandée. S'il est nécessaire, cliquez sur le menu Tools ⇒ Add-ins et cochez la case Analysis Toolpak et la case Solver add-in qui servira un peu plus tard, par la même occasion. Recommencez l'étape précédente.

⇒ Cliquez sur le bouton d'accès à la feuille correspondant à Input Y Range et sélectionnez la plage de nom PRICE qui contient les prix des maisons et l'en-tête.



- ⇒ Faites de même pour Input X Range et sélectionnez la plage de nom HABSIT qui contient les superficies habitables et les situations, avec les en-têtes.
- ⇒ Comme les en-têtes sont sur la ligne avant les données, cochez la case Labels.



- ⇒ Cliquez sur le bouton New Workbook. Cochez les cases relatives aux résidus et aux graphiques (*plots*), pour les besoins futurs. Cliquez sur OK.

### Remarques



1. Une importante restriction de l'outil Regression de la boîte à outils complémentaire Analysis Toolpack est la suivante : il faut que la plage contenant les valeurs de l'ensemble des variables explicatives soit de forme rectangulaire. Il n'est donc pas question de fournir deux colonnes qui ne commencent pas sur la même ligne ni deux colonnes qui ne sont pas juxtaposées.
2. Il est possible de réaliser un ajustement sans constante ou ajustement par l'origine. Il suffit pour cela de cocher la case Constant is zero. Ce n'est toutefois que rarement recommandé.
3. Nous avons remarqué une erreur étrange de l'outil Regression d'Excel 97 qui refuse de fonctionner quand il y a des cellules fusionnées. Pour cette raison, les liens de la feuille ne sont actifs que dans la première cellule.



### 1.3 INTERPRÉTATION DES RÉSULTATS

⇒ Un classeur est créé. Sauvez-le, dans le dossier où vous mettez les travaux du chapitre 7, sous le nom CH07EX01Regression.XLS, par exemple.

? Vérifiez que les informations qui apparaissent dans le nouveau classeur sont celles qui sont données en annexe à la fin de la partie 1. Est-ce correct ?



#### 1.3.1 Votre réponse

La sortie comporte six rubriques :

1. *Regression Statistics*.
2. ANOVA. Nous commenterons ce tableau dans la partie 5.
3. *Coefficients*. Limitons-nous aux coefficients estimés — pour le restant, voir les parties 2, 5 ou 6 — la constante  $b_0$  sur la ligne Intercept, le coefficient de régression  $b_1$  sur la ligne HAB et le coefficient de régression  $b_2$  sur la ligne SIT.
4. *Residual Output*. Voir la partie 6.
5. *Probability Output*. Voir la partie 6.
6. Les graphiques. Voir la partie 6.

? Donnez la valeur des coefficients estimés et écrivez l'équation de régression.



#### 1.3.2 Vos réponses

## SYNTHÈSE

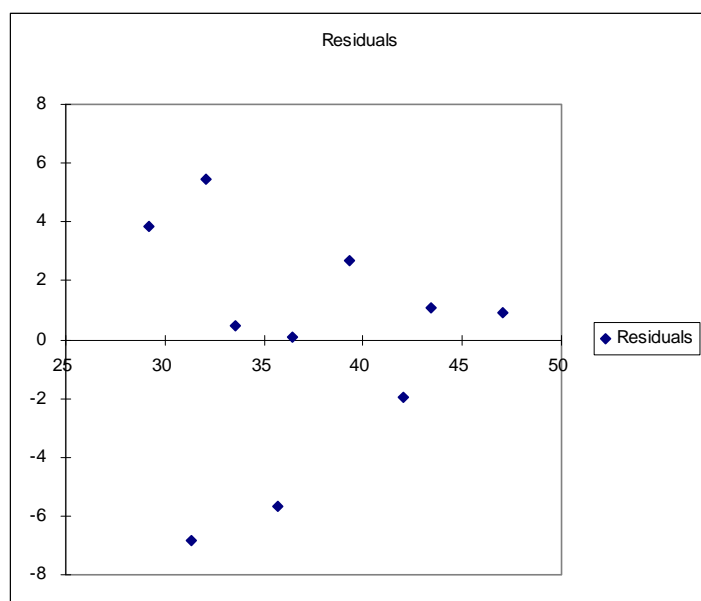
Nous avons vu — ou rappelé — comment employer la régression linéaire multiple dans Excel. La méthode est la même que celle employée au chapitre 2. Nous irons plus loin dans l'interprétation des résultats.



## Annexe. Les résultats de régression

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.8309378					
R Square	0.6904576					
Adjusted R Square	0.602017					
Standard Error	4.4210168					
Observations	10					
ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	2	305.1822733	152.59	7.80701	0.016501383	
Residual	7	136.8177267	19.545			
Total	9	442				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>tStat</i>	<i>P-value</i>	<i>Lower95%</i>	<i>Upper95%</i>
Intercept	18.283569	7.683312583	2.3796	0.04891	0.115434637	36.4517
HAB	0.725265	0.347654593	2.0862	0.0754	-0.096806877	1.547337
SIT	9.2020024	3.090952765	2.9771	0.0206	1.893065716	16.51094

RESIDUAL OUTPUT		
<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>
1	29.1625442	3.83745583
2	35.6899293	-5.689929329
3	32.0636042	5.43639576
4	39.3162544	2.683745583
5	43.4414016	1.058598351
6	33.5141343	0.485865724
7	41.9908716	-1.990871614
8	31.3383392	-6.838339223
9	47.0677267	0.932273263
10	36.4151943	0.084805654





**Partie 2** Dans cette partie, le but de l'exercice est d'introduire les mesures de qualité de l'ajustement que sont la somme des carrés des résidus et le coefficient de détermination  $R^2$ . Nous commençons par regarder les résultats obtenus pour le modèle de la partie 1. Ensuite, nous comparons les résultats avec ceux d'un modèle avec HAB comme unique variable explicative.

### 2.1 LE COEFFICIENT DE DÉTERMINATION

Le coefficient de détermination représente, comme dans la régression linéaire simple, la proportion de la variance de la variable dépendante, PRICE, qui est expliquée par l'ensemble des variables explicatives, ici HAB et SIT. Regardez, au choix, les résultats dans le classeur CH07EX01Regression.XLS ou dans le tableau en annexe de la partie 1.



Que vaut la mesure  $R^2$  de qualité de l'ajustement appelée R square ? Interprétez cette mesure.



2.1.1 Votre réponse

### 2.2 LA SOMME DES CARRÉS RÉSIDUELLE

La somme des carrés résiduelle est présentée dans le tableau ANOVA, sur la ligne Residual et dans la colonne intitulée SS. La variance résiduelle est sur la même ligne, dans la colonne intitulée MS.



Que vaut la somme des carrés résiduelle ? Notez également la variance résiduelle.



2.2.1 Votre réponse

### 2.3 COMPARAISON AVEC LA RÉGRESSION LINÉAIRE SIMPLE

Suivez les instructions de la partie 1 pour estimer les paramètres d'un modèle de régression linéaire simple pour la variable PRICE avec la superficie habitable comme seule variable explicative. Vous trouverez les résul-



tats déjà complétés en cliquant sur l'onglet PriceHab du classeur.



?

Donnez la valeur des coefficients estimés et écrivez l'équation de régression.

*2.3.1 Vos réponses*



?

Que vaut le coefficient de détermination  $R^2$  et interprétez le résultat obtenu ? Comparez avec ce qui a été obtenu pour la régression multiple.

*2.3.2 Vos réponses*



?

Que vaut la somme des carrés résiduelle ? Comparez avec ce qui a été obtenu pour la régression multiple. Notez également la variance résiduelle.

*2.3.3 Vos réponses*

## SYNTHÈSE

Nous avons examiné le coefficient de détermination et la somme des carrés résiduelle pour le modèle de la partie 1. Nous avons aussi comparé les résultats avec ceux d'un modèle de régression linéaire simple avec la superficie habitable HAB comme seule variable explicative.



**Partie 3** Le but de cette partie de l'exercice est d'employer un logiciel conçu pour le traitement de données temporelles. Ce n'est pas un logiciel conçu spécifiquement pour la régression multiple mais il offre néanmoins un peu plus de souplesse qu'Excel.

### 3.1 INTRODUCTION

Nous allons effectuer cette étude en employant Time Series Expert for Windows, en abrégé TSE.

- ⇒ Suivez les instructions rappelées en annexe du document introductif du cours afin de lancer le logiciel.
- ⇒ Choisissez le répertoire de données approprié sur votre disque (pas sur le CD-ROM) : menu File ⇒ Open. Choisissez DATA puis CHAP07 puis CH07EX01.
- ⇒ Chargez le problème déjà préparé : HABITAT1. Vous devez alors voir dans le bas de l'écran que la variable dépendante est PRICE, que l'échantillon d'estimation est 1 – 10 et que les prévisions seront calculées jusqu'en 10.
- ⇒ Pour visualiser le tableau des données : menu Data ⇒ Spreadsheet.
- ⇒ Pressez la touche fonction F3 pour charger une série dans la colonne A du tableau. Sélectionnez PRICE.
- ⇒ Avec la flèche à droite, placez le curseur dans la colonne B.
- ⇒ Pressez la touche fonction F3 pour charger une série dans la colonne B du tableau. Sélectionnez HAB.
- ⇒ Avec la flèche à droite, placez le curseur dans la colonne C.
- ⇒ Pressez la touche fonction F3 pour charger une série dans la colonne C du tableau. Sélectionnez SIT.
- ⇒ Quittez le tableur par le menu File ⇒ Exit TSE Spreadsheet.



- ⇒ Pour visualiser graphiquement les séries : menu Graphics ⇒ Series.
- ⇒ Cliquez Select. Sélectionnez PRICE puis cliquez Open.
- ⇒ Cliquez OK pour obtenir le graphique.



### Remarques

1. Pour sélectionner plusieurs variable pour un graphique, sélectionnez la première, enfoncez la touche Ctrl et sélectionnez les suivantes enfin relâchez la touche Ctrl et cliquez Open puis OK pour valider.
2. Dans tous les graphiques de TSE, pour quitter le mode graphique, fermez la fenêtre.

## 3.2 ESTIMATION DE LA RÉGRESSION SIMPLE EN FONCTION DE HAB

Une variable explicative a déjà été définie dans le problème en plus de la constante. Il s'agit de HAB. Nous allons maintenant estimer les paramètres du modèle.

- ⇒ Pour estimer les paramètres d'une régression linéaire, menu Methods ⇒ Ordinary least squares ⇒ Expertise.
- ⇒ Dans la fenêtre de dialogue, sur la ligne Save residuals, vous devez voir "RES". Sur la ligne Save forecasts, vous devez voir "PRED". Cliquez OK. Sauvegardez le rapport.

Au début de la sortie, on voit que la liste des variables explicatives contient la CONSTANTE ainsi que HAB.



Ecrivez l'équation du modèle de régression et comparez-la à la réponse 2.3.1 donnée dans la partie 2.



3.2.1 Votre réponse



**?**

Que vaut la valeur de  $R^2$  sur la ligne R square ? Comparez avec la réponse 2.3.2.

*3.2.2 Votre réponse*

**?**

Que vaut la quantité sur la ligne Residual variance ? Comparez avec la réponse 2.3.2. Notez également la quantité sur la ligne Residual standard deviation.

*3.2.3 Vos réponses*

### 3.3 ESTIMATION DE LA RÉGRESSION MULTIPLE

Nous allons maintenant charger un problème dans lequel les variables explicatives sont HAB et SIT.

- ⇒ Chargez le problème déjà préparé : HABITAT2.
- ⇒ Pour estimer les paramètres d'une régression linéaire, menu Methods ⇒ Ordinary least squares ⇒ Expertise.
- ⇒ Dans la fenêtre de dialogue, sur la ligne Save residuals, vous devez voir "RES". Sur la ligne Save forecasts, vous devez voir "PRED". Cliquez OK. Sauvegardez le rapport.

Au début de la sortie, on voit que la liste des variables explicatives contient la CONSTANTE ainsi que HAB et SIT.

**?**

Ecrivez l'équation du modèle de régression et comparez-la à la réponse 1.3.2 donnée dans la partie 1.





### 3.3.1 Votre réponse



Que vaut la valeur de  $R^2$  sur la ligne R square ? Comparez avec la réponse 2.1.1 donnée dans la partie 2. Interprétez-la.



### 3.3.2 Votre réponse



Que vaut la quantité sur la ligne Residual variance ? Comparez avec la réponse 2.2.1. Notez également la quantité sur la ligne Residual standard deviation.



### 3.3.3 Vos réponses

## 3.4 EXAMEN DES VALEURS PRÉDITES ET DES RÉSIDUS



Pour obtenir le graphique des valeurs ajustées et prévisions en parallèle avec les données, procédez comme suit : menu Graphics ⇒ Predictions/Forecasts. Cliquez OK.

Comme pour tous les graphiques de TSE, il est possible de focaliser sur un point.

Un graphique révélateur de la qualité de l'ajustement est constitué des résidus en fonction des valeurs ajustées. C'est le *diagramme des résidus*.



Pour l'obtenir, procédez comme suit : menu Graphics ⇒ Xy graphic/Scatter diagram



Cliquez Select Y et sélectionnez RES puis cliquez Open.



Cliquez Select X et sélectionnez PRED puis cliquez Open.





Cliquez OK pour obtenir le graphique.



Examinez la présence éventuelle d'une structure dans ce graphique.

3.4.1 Votre réponse

*Facultatif*

### 3.5 CONSTRUCTION DE VARIABLES SUPPLÉMENTAIRES

Nous définissons maintenant plusieurs variables supplémentaires relativement artificielles qui vont servir dans les illustrations. Il s'agit des variables suivantes :

- HABSIT définie comme le produit  $HAB \cdot SIT$
- HAB2 définie comme le carré de HAB :  $HAB^2$
- HAB3 définie comme le cube de HAB :  $HAB^3$

et d'autres encore. Vous pouvez sauter cette étape si vous le souhaitez. Dans ce cas, vous ne pourrez pas effectuer complètement le paragraphe 3.6. mais vous pourrez néanmoins vous baser sur les fichiers de sortie fournis.



Pour effectuer ces transformations : menu Data  $\Rightarrow$  Spreadsheet.



Pressez la touche fonction F3 pour charger une série dans la colonne A du tableau. Sélectionnez HAB et cliquez Open.



Cliquez dans la colonne B. Pressez la touche fonction F3 pour charger une série dans la colonne B du tableau. Sélectionnez SIT et cliquez Open.



Cliquez dans la colonne C. Utilisez le menu Operation du tableur et cliquez Enter Oper pour effectuer une opération entre les colonnes du tableau. Cliquez A dans la colonne Operand 1 puis \* puis B dans la colonne Operand 2 et cliquez OK. Le produit  $HAB \cdot SIT$  est calculé dans la colonne C. Pressez la touche F10 pour renommer la série. Tapez HABSIT. Pressez la touche F2 pour sauver la série.



- ⇒ Cliquez dans la colonne D. Utilisez le menu Operation du tableur et cliquez Enter Oper pour effectuer une opération entre les colonnes du tableau. Cliquez A dans la colonne Operand 1 puis \* puis A dans la colonne Operand 2 et cliquez OK. Le carré de HAB est calculé dans la colonne D. Pressez la touche F10 pour renommer la série. Tapez HAB2. Pressez la touche F2 pour sauver la série.
- ⇒ Quittez le tableur par menu File ⇒ Exit TSE Spreadsheet.

### 3.6 UTILISATION DE VARIABLES SUPPLÉMENTAIRES

*Facultatif*

Nous allons maintenant considérer la régression de PRICE en fonction de HAB, SIT et du produit HABSIT. Vous pouvez sauter cette étape si vous le souhaitez et aller au paragraphe 3.7.

- ⇒ Chargez le problème déjà préparé : HABITAT3.
- ⇒ Pour estimer les paramètres d'une régression linéaire, menu Methods ⇒ Ordinary least squares ⇒ Expertise.
- ⇒ Dans la fenêtre de dialogue, sur la ligne Save residuals, vous devez voir "RES". Sur la ligne Save forecasts, vous devez voir "PRED". Cliquez OK. Sauvegardez le rapport.

?

(a) Ecrivez l'équation du modèle de régression et notez (b) la valeur de  $R^2$  sur la ligne R square, (c) la quantité sur la ligne Residual standard deviation ainsi que (d) le nombre mentionné sur la ligne Adjusted R square.

#### 3.6.1 Vos réponses

- (a)
- (b)
- (c)
- (d)



- ⇒ Faites de même pour la régression de PRICE en fonction de HAB, SIT, du produit HABSIT et de  $HAB^2$ . Le problème s'appelle



## HABITAT4.

?

(a) Écrivez l'équation du modèle de régression et notez (b) la valeur de  $R^2$  sur la ligne R square, (c) la quantité sur la ligne Residual standard deviation ainsi que (d) le nombre mentionné sur la ligne Adjusted R square.



## 3.6.2 Vos réponses

(a)

(b)

(c)

(d)

## 3.7 SYNTHÈSE DES RÉSULTATS

Si vous avez effectué entièrement les deux paragraphes précédents, vous avez déjà les réponses 3.7.1 et 3.7.2. Dans le cas contraire, veuillez ouvrir les fichiers de sortie préparés pour recueillir les informations demandées.



Si nécessaire, chargez le fichier de résultats pour la régression de PRICE en fonction de HAB, SIT et HABSIT : menu File  $\Rightarrow$  Report  $\Rightarrow$  All. Choisissez HABITAT3. Sauvegardez le rapport.

?

(a) Ecrivez l'équation du modèle de régression et notez (b) la valeur de  $R^2$  sur la ligne R square, (c) la quantité sur la ligne Residual standard deviation ainsi que (d) le nombre mentionné sur la ligne Adjusted R square.



## 3.7.1 Votre réponse

(a)



(b)

(c)

(d)



Si nécessaire, chargez le fichier de résultats pour la régression de PRICE en fonction de HAB, SIT, HABSIT et  $HAB^2$  : menu Reports  $\Rightarrow$  Report. Choisissez HABITAT4.



(a) Ecrivez l'équation du modèle de régression et notez (b) la valeur de  $R^2$  sur la ligne R square, (c) les quantités sur la ligne Residual standard deviation et (d) Adjusted R square.



### 3.7.2 Votre réponse

(a)

(b)

(c)

(d)

Faites de même pour les rapports suivants.



HABITAT5 ?



### 3.7.3 Votre réponse

(a)

(b)

(c)

(d)



**?****HABITAT6 ?***3.7.4 Votre réponse*

(a)

(b)

(c)

(d)

**?****HABITAT7 ?***3.7.5 Votre réponse*

(a)

(b)

(c)

(d)

**?**

Quel est le meilleur modèle au sens de celui qui a le plus petit  $R^2$  ? Ceci vous semble-t-il logique ?

*3.7.6 Votre réponse***Remarques**

1. La formule suivante, vue au chapitre 2, reste valable :





$$\text{MSE} = s_y^2 (1 - R^2).$$

2. Nous appelons variance résiduelle, notée  $\hat{\sigma}^2$ , la quantité  $n\text{MSE}/(n - k)$ . C'est une estimation de la variance  $\sigma^2$  des erreurs du modèle. Sa racine carrée est l'écart-type résiduel, noté  $\hat{\sigma}$ . Voir la justification dans la partie suivante de l'exercice.

## SYNTHÈSE

En employant le logiciel Time Series Expert, nous avons retrouvé les résultats précédents. Nous avons pu aisément définir de nouvelles variables et avons effectué plusieurs estimations de modèle de régression. Nous avons procédé à l'interprétation du coefficient de détermination  $R^2$ . Nous avons pu constater que le coefficient de détermination augmente, ou reste au moins constant, quand on introduit une variable explicative supplémentaire. Nous envisagerons une solution à ce problème dans la partie 4 de l'exercice.



**Partie 4** Dans cette partie de l'exercice, le but est d'introduire les mesures de qualité de l'ajustement que sont le coefficient de détermination  $R^2$  corrigé ou  $\bar{R}^2$  ainsi que la variance résiduelle et l'écart-type résiduel.

L'essentiel du travail de cette partie de l'exercice a été réalisé dans la partie 3 de l'exercice. Nous allons ici partir du tableau qui a été réalisé et le compléter par un autre tableau.

#### 4.1 RÉSULTATS POUR LE COEFFICIENT DE DÉTERMINATION

Si nous résumons les résultats des modèles de la partie 3 relatifs au coefficient de détermination et à la somme des carrés résiduelle, nous obtenons le tableau suivant :

Variables explicatives	$R^2$	Somme des carrés résiduelle
HAB	0,299	310
HAB, SIT	0,690	137
HAB, SIT, HAB.SIT	0,698	133
HAB, SIT, HAB.SIT, $HAB^2$	0,741	114
HAB, SIT, HAB.SIT, $HAB^2$ , $HAB^3$	0,742	114
HAB, SIT, HAB.SIT, $HAB^2$ , $HAB^3$ , $HAB^4$	0,791	93
HAB, SIT, HAB.SIT, $HAB^2$ , $HAB^3$ , $HAB^4$ , $HAB^5$	0,852	65
HAB, SIT, HAB.SIT, $HAB^2$ , $HAB^3$ , $HAB^4$ , $HAB^5$ , $HAB^6$	(*)	(*)

À titre d'exercice, on peut retrouver ces résultats dans Excel et compléter le classeur CH07EX01Regression.XLS. Par exemple, on peut effectuer l'estimation du premier modèle avec HAB pour seule variable explicative et sauver les résultats dans une feuille par exemple PriceHab. Ces résultats ont été ajoutés dans la feuille PriceHab du classeur CH07EX01.XLS.

Posons à nouveau la dernière question de la partie 3.



Quel est le meilleur modèle au sens de celui qui a le plus petit  $R^2$  ? Ceci vous semble-t-il logique ?





## 4.1.1 Votre réponse

## 4.2 DÉFINITION DU COEFFICIENT DE DÉTERMINATION CORRIGÉ

Nous avons également noté le coefficient de détermination corrigé « adjusted R square » souvent noté  $\bar{R}^2$ . Il est défini de la manière suivante :

$$\bar{R}^2 = R^2 - (1 - R^2) \frac{k-1}{n-k}.$$

**Remarques**

1.  $\bar{R}^2 < R^2$  et  $\bar{R}^2$  peut même être négatif, contrairement à  $R^2$ .
2. L'introduction du coefficient de détermination corrigé se justifie en considérant la régression dans le contexte d'une population plutôt que de celui d'un échantillon issu de celle-ci. Au lieu des variances d'échantillon, comportant le nombre d'observations  $n$  au dénominateur, on est alors tenté d'employer des estimations non biaisées des variances dans la population. Ces estimations sont donc corrigées pour le biais en tenant compte des nombres de degrés de liberté adéquats. Pour la variance d'y,  $s_y^2$ , il suffit d'employer  $n - 1$  au lieu de  $n$ . Pour la variance résiduelle,  $\hat{\sigma}^2$ , obtenue en estimant les  $k$  coefficients de régression, le nombre de degré de liberté est  $n - k$  :  $\hat{\sigma}^2 = n\text{MSE}/(n - k)$ . On peut vérifier que la formule de  $\bar{R}^2$  est la solution, exprimée en termes de  $R^2$ , d'une équation similaire à celle trouvée pour  $R^2$ , en remplaçant les deux variances par des estimations non biaisées :

$$\frac{n}{n-k} \text{MSE} = \frac{n}{n-1} s_y^2 (1 - \bar{R}^2).$$

Par exemple, considérons les résultats donnés en annexe de la partie 1 et discutés dans la partie 2, pour les prix de ventes d'habitations en fonction de HAB et SIT, où  $R^2 = 0,6905$ .

*facultatif*



Vérifiez la valeur de  $\bar{R}^2$  dans cet exemple

## 4.2.1 Votre réponse





Vérifiez la valeur de  $\bar{R}^2$  dans cet exemple en employant l'approche alternative exprimée dans la remarque 2.

*facultatif*



#### 4.2.2 Votre réponse

### 4.3 RÉSULTATS POUR LE COEFFICIENT DE DÉTERMINATION CORRIGÉ

Nous pouvons maintenant reprendre l'exercice qui avait consisté à introduire comme variables explicatives des variables telles que le produit HAB.SIT et les puissances successives de HAB. Le tableau suivant contient les valeurs de  $\bar{R}^2$  et de l'écart-type résiduel  $\hat{\sigma}$ , en plus de celles de  $R^2$ .

Variables explicatives	$R^2$	$\bar{R}^2$	$\hat{\sigma}$
HAB	0,299	0,211	6,23
HAB, SIT	0,690	0,602	4,42
HAB, SIT, HAB.SIT	0,698	0,547	4,72
HAB, SIT, HAB.SIT, HAB <sup>2</sup>	0,741	0,535	4,78
HAB, SIT, HAB.SIT, HAB <sup>2</sup> , HAB <sup>3</sup>	0,742	0,420	5,34
HAB, SIT, HAB.SIT, HAB <sup>2</sup> , HAB <sup>3</sup> , HAB <sup>4</sup>	0,791	0,372	5,55
HAB, SIT, HAB.SIT, HAB <sup>2</sup> , HAB <sup>3</sup> , HAB <sup>4</sup> , HAB <sup>5</sup>	0,852	0,334	5,71
HAB, SIT, HAB.SIT, HAB <sup>2</sup> , HAB <sup>3</sup> , HAB <sup>4</sup> , HAB <sup>5</sup> , HAB <sup>6</sup>	(*)	(*)	(*)



Quel est le meilleur modèle au sens de celui qui a le plus petit  $\bar{R}^2$  ? Ceci vous semble-t-il logique ?

#### 4.3.1 Votre réponse



Quel est le meilleur modèle au sens de celui qui a la plus petite variance résiduelle  $\hat{\sigma}^2$  ou le plus petit écart-type résiduel  $\hat{\sigma}$  ? Ceci



vous semble-t-il logique ?

4.3.2 Votre réponse



## SYNTHÈSE

Nous avons vu que le coefficient de détermination n'est pas une mesure adéquate de la qualité d'un modèle dans le contexte où le nombre de variables explicatives est susceptible de varier. À sa place, nous avons proposé le coefficient de détermination corrigé et la variance résiduelle ou l'écart-type résiduel.



**Partie 5** Dans cette partie de l'exercice, les principaux aspects d'inférence statistique sont examinés : tests sur les coefficients, test global sur le modèle. Nous nous référons à l'exemple traité dans la section 1 où PRICE est la variable dépendante et HAB et SIT sont deux variables explicatives dans un modèle de régression multiple avec constante. En particulier, nous utilisons un extrait du tableau en annexe de la partie 1 que nous reprenons ici pour la commodité :

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.8309378					
R Square	0.69045763					
Adjusted R Square	0.60201696					
Standard Error	4.4210168					
Observations	10					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	2	305.1822733	152.59	7.80701	0.016501383	
Residual	7	136.8177267	19.545			
Total	9	442				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	18.2835689	7.683312583	2.3796	0.04891	0.115434637	36.4517
HAB	0.72526502	0.347654593	2.0862	0.0754	-0.096806877	1.547337
SIT	9.20200236	3.090952765	2.9771	0.0206	1.893065716	16.51094

Supposons que les données soient relatives à un échantillon prélevé dans une population, de façon aléatoire ("*random sample*"). On ne s'intéresse pas aux coefficients de régression pour l'échantillon mais bien à ceux qui sont valables pour la population entière dont cet échantillon est issu. On distingue  $b_j$ , le vrai coefficient, celui de la population et qui est inconnu, et  $\hat{b}_j$ , son estimation dans l'échantillon, obtenue par les formules données plus haut.  $\hat{b}_j$  dépendant de l'échantillon, il est considéré comme une variable aléatoire. Moyennant certaines suppositions concernant la population (voir l'exercice 4), la distribution de  $\hat{b}_j$  peut être déterminée.

### 5.1 L'ERREUR TYPE, UNE MESURE D'INCERTITUDE

À chaque coefficient dont la valeur est reprise dans la colonne « Coefficients », correspond une erreur type dans la colonne « Standard error » et une statistique  $t$  ou statistique de Student, dans la colonne « tStat ».

L'erreur-type associée à un coefficient  $\hat{b}_j$  qu'on notera souvent  $\hat{\sigma}(\hat{b}_j)$ , à lire *écart-type estimé de l'estimation*  $\hat{b}_j$ , apparaît comme mesurant l'incertitude sur le coefficient estimé correspondant.

L'erreur-type est une estimation de la dispersion du coefficient estimé par rapport aux variations d'échantillonnage. En effet, les résultats présentés ici sont relatifs à un échantillon déterminé. D'autres échantillons donne-



raient lieu à d'autres estimations des coefficients. C'est cette variabilité que représente l'erreur-type. Nous verrons comment on peut l'obtenir dans la partie D de l'exercice avancé, d'autant que l'erreur-type sera déterminée en employant un seul échantillon.

L'erreur-type est exprimée dans les mêmes unités que les coefficients. Par exemple le coefficient de HAB est en milliers d'euros par m<sup>2</sup>, donc également l'erreur-type correspondante.



?

Quelle est l'erreur-type correspondant à la constante ?

5.1.1 Votre réponse



?

Quelle est l'erreur-type correspondant au coefficient de la variable HAB ?

5.1.2 Votre réponse



?

Quelle est l'erreur-type correspondant au coefficient de la variable SIT ?

5.1.3 Votre réponse



### Remarque

Au lieu d'employer l'outil de régression pour obtenir tous ces résultats, une autre solution dans Excel consiste à employer la fonction "LINEST" ("DROITEREG" dans la version française, "LIJNSCH" dans la version néerlandaise) qui renvoie tous les résultats mentionnés ici. Cette solution est automatique mais est moins simple dans la mesure où les résultats sont présentés, presque en vrac dans une matrice à 4 lignes et qu'il faut de sérieuses connaissances pour récupérer les informations pertinentes.

## 5.2 TEST D'HYPOTHÈSE SUR UN COEFFICIENT DU MODÈLE





Dans les sorties de la régression de PRICE en fonction de HAB et de SIT, vérifiez que les statistiques  $t$  ou statistiques de Student sont égales aux rapports entre le coefficient et l'erreur-type (notée standard-error) ?



#### 5.2.1 Votre réponse

Dans la présentation des résultats, on accompagne fréquemment les coefficients estimés soit de leur erreur-type (généralement précédée de  $\pm$ ), soit de leur statistique  $t$  avec son signe éventuel. Par exemple, on écrira l'équation de régression de la réponse 1.3.2 sous la forme :

$$\text{PRICE} = 18,28 + 0,725 \text{ HAB} + 9,20 \text{ SIT} \\ (\pm 7,68) (\pm 0,348) (\pm 3,09)$$

ou

$$\text{PRICE} = 18,28 + 0,725 \text{ HAB} + 9,20 \text{ SIT} \\ (2,38) (2,09) (2,98)$$

L'idée est d'employer la statistique de Student pour tester l'hypothèse que le coefficient correspondant *dans la population* est égal à 0 (hypothèse nulle).

#### Remarque

Nous verrons dans l'exercice 2 une justification empirique de ce qui suit.



Que vaut la statistique de test pour HAB ? Expliquer de quelle hypothèse il s'agit ? Comment pourrait-on l'interpréter ?



#### 5.2.2 Votre réponse

Sous certaines conditions qui sont données dans le cours et détaillées dans



l'exercice 4, et *en supposant que l'hypothèse est vraie*, la distribution de la statistique  $t$  est la distribution de Student dont le nombre de degrés de liberté est le nombre d'observations diminué du nombre de paramètres estimés. Ici, nous avons 10 observations et 3 paramètres estimés. On a donc une distribution de Student à 7 degrés de liberté. Pour la commodité, nous emploierons souvent l'approximation de la distribution de Student par une distribution normale centrée réduite. On remarque la similitude avec l'inférence statistique sur la moyenne. La distinction entre test unilatéral et test bilatéral s'applique également ici. En général, nous effectuerons des tests bilatéraux.

On spécifie un *niveau de signification* ("level of significance"), une probabilité que nous prendrons ici égale à 5%. Ce niveau ou seuil de signification ou niveau ou seuil du test représente la probabilité de *rejeter l'hypothèse* à partir de l'échantillon alors qu'en fait cette hypothèse est vraie dans la population.

Rappelons que le quantile d'ordre  $0,975 = 1 - 0,05/2$  d'une distribution normale centrée réduite vaut à peu près 1,96. La distribution normale centrée réduite étant symétrique autour de 0, le quantile d'ordre  $0,025 = 0,05/2$  vaut l'opposé,  $-1,96$ .

Si l'hypothèse est vraie, on a donc approximativement une probabilité égale à  $0,975 - 0,025 = 0,95$  que  $t$  tombe dans l'intervalle  $[-1,96 ; 1,96]$ .

Au niveau de 5%, et en employant l'approximation normale pour simplifier, on rejette l'hypothèse si  $t < -1,96$  ou si  $t > 1,96$ . On dit alors que le coefficient  $\hat{b}_j$ , calculé sur l'échantillon, est *significatif* ("significant") au niveau ou au seuil de 5%. Par conséquent, 5 fois sur 100, on court le risque de rejeter une hypothèse qui est vraie, en réalité. Dans le cas où  $-1,96 < t < 1,96$ , on peut accepter provisoirement l'hypothèse que  $b_j$  soit nul (dans la population) et on dit que  $\hat{b}_j$  (dans l'échantillon) est *non significatif* ("un-significant") au niveau de 5%.

### Remarque

Les tests effectués sont dits *bilatéraux* ("two-tail" ou "two-sided"), c'est-à-dire qu'on rejette l'hypothèse à tester, que le coefficient soit positif ou négatif et de valeur absolue trop grande. Quand on a une idée *a priori* sur le signe du coefficient dans la population, on peut effectuer un test *unilatéral* ("one-tail" ou "one-sided").





**?**

Rejette-t-on l'hypothèse que le coefficient de régression de HAB égal à 0, au niveau de 5% ?

### 5.2.3 Votre réponse

Plutôt que 1,96 ou 2, il est préférable d'employer le quantile d'ordre 0,975 de la loi de Student à  $n - k$  degrés de liberté, surtout quand le nombre de données  $n$  est petit,. Ces quantiles figurent, pour quelques valeurs sélectionnées de  $\alpha$ , dans les tables de la plupart des ouvrages de statistique, notamment Moore et McCabe (1998) et Wonnacott et Wonnacott (1991).

### Remarque



Pour trouver cette valeur critique dans Excel, entrez la formule suivante dans une cellule : « =TINV(0.05;7) ».

La distribution peut d'autant mieux être approchée par une normale centrée réduite que le nombre de degrés de liberté est grand, avec une bonne approximation quand  $n - k$  est supérieur à 30, disons.

**?**

Sachant que le quantile d'ordre 97,5% d'une distribution de Student à 7 degrés de liberté vaut à peu près 2,364, rejette-t-on l'hypothèse que le coefficient de HAB égal à 0, au niveau de 5% ?



### 5.2.4 Votre réponse

**?**

Effectuez de manière similaire le test pour la constante, d'une part en employant la distribution approchée centrée normale réduite, d'autre part, en employant la distribution exacte de Student à 7 degrés de liberté, dans les deux cas un seuil de 5 % ?



### 5.2.5 Votre réponse





?

Effectuez de manière similaire le test pour le coefficient de SIT, d'une part en employant la distribution approchée centrée normale réduite, d'autre part, en employant la distribution exacte de Student à 7 degrés de liberté, dans les deux cas un seuil de 5 % ?

5.2.6 Votre réponse

Par ailleurs, beaucoup de logiciels statistiques fournissent, à côté du coefficient estimé, de son erreur-type et de la statistique  $t$ , la *probabilité de signification*  $P$  ou *valeur*  $P$  associée qui représente le niveau de signification maximal pour lequel le coefficient serait significatif.

Il est possible d'utiliser cette information pour effectuer un test au niveau de probabilité quelconque  $\alpha$ . Si  $P$  est inférieur au  $\alpha$  choisi, le coefficient de régression est significatif, tandis que si  $P$  est supérieur à  $\alpha$ , le coefficient est non significatif.



?

Recherchez les valeurs  $P$  dans le tableau des sorties d'Excel correspondant à chacun des coefficients. Utilisez-les pour effectuer les tests au niveau de 5% au lieu d'employer les statistiques  $t$  ?

5.2.7 Votre réponse



?

Les conclusions sont-elles identiques aux conclusions obtenues en comparant les statistiques  $t$  à 1,96, approximation provenant de la distribution centrée normale réduite ?

5.2.8 Votre réponse





Les conclusions sont-elles identiques aux conclusions obtenues en comparant les statistiques  $t$  à 2,364, provenant de la distribution exacte de Student à 7 degrés de liberté ?



5.2.9 Votre réponse

### Remarque



Plus généralement, on pourrait effectuer un test de l'hypothèse que le coefficient  $b_j$ , le vrai coefficient population, est égal à un nombre déterminé  $b_j^{(0)}$ , au lieu de 0. Dans ce cas, la statistique de Student doit prendre la forme  $\frac{\hat{b}_j - b_j^{(0)}}{\hat{\sigma}(\hat{b}_j)}$  mais la distribution sous l'hypothèse que  $b_j = b_j^{(0)}$  est également une distribution de Student à  $n - k$  degrés de liberté, où  $n$  est le nombre d'observations et  $k$  est le nombre de paramètres estimés.

## 5.3 TEST D'HYPOTHÈSE SUR LE MODÈLE

L'exposé qui précède est relatif à un coefficient de régression quelconque mais spécifique. Il est fréquent de s'interroger sur la validité globale du modèle : le modèle est-il globalement explicatif ? Cela revient à tester l'hypothèse que tous les coefficients de régression sont simultanément égaux à zéro, en espérant une réponse négative qu'il faudrait interpréter comme suit : au moins une des variables explicatives contribue à expliquer la variable dépendante d'une manière qui laisse peu de place au hasard. En pratique, il y a souvent une constante dans le modèle et celle-ci n'est pas considérée comme explicative (d'ailleurs, le cas où  $R^2 = 0$  implique que la meilleure prédiction est fournie par la moyenne). Par conséquent, on se contente de tester l'hypothèse que les coefficients des variables explicatives effectives soient égaux à zéro. Pour le modèle d'équation

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{k-1}x_{k-1} + e$$

on teste l'hypothèse  $b_1 = b_2 = \dots = b_{k-1} = 0$  (mais pas  $b_0 = 0$ ).

La statistique de test est la statistique  $F$  présentée dans le tableau ANOVA. Il est possible de montrer que, sous les mêmes suppositions que celles utilisées pour les statistiques de Student (voir exercice 4), et à condition que l'hypothèse soit vraie, le rapport suivant, appelé statistique  $F$  a une distribution de Fisher-Snedecor avec  $k - 1$  degrés de liberté au numérateur et  $n$



–  $k$  degrés de liberté au dénominateur.



?

Quelle est la valeur de la statistique  $F$  ? Quels sont les degrés de liberté du numérateur et du dénominateur ?

5.3.1 Votre réponse



?

Sachant que le quantile d'ordre 95 % d'une distribution de Fisher à 2 et 7 degrés de liberté vaut 4,74, est-ce que l'on rejette l'hypothèse que les coefficients des deux variables HAB et SIT soient simultanément égaux à 0 au niveau de 5 % ? Peut-on donc considérer que le modèle est valable ?

5.3.2 Votre réponse



### Remarques

1. Pour trouver cette valeur critique dans Excel, entrez la formule suivante dans une cellule : « =FINV(0.05;2 ;7) ».

2. La statistique  $F$  pour le test de nullité de tous les coefficients, sauf la constante, peut être évaluée à partir des autres résultats imprimés. En effet, on peut l'obtenir par le rapport que voici :

$$\frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}}$$

Cette quantité peut être calculée dès que le coefficient de détermination est connu.

## 5.4 CONSTRUCTION D'UN INTERVALLE DE CONFIANCE

En plus des tests d'hypothèses, les méthodes statistiques proposent des intervalles de confiance. Un intervalle de confiance pour un paramètre de la population est un intervalle déterminé à l'aide de l'échantillon qui a une probabilité donnée de contenir la vraie valeur du paramètre de la popula-



tion. On se fixe une probabilité appelée *probabilité de confiance* ("confidence probability"). Notons la  $1 - \alpha$ . Quand on connaît la loi de la statistique étudiée, exacte ou approchée, et qu'elle dépend exclusivement du paramètre en question, on peut essayer de déterminer les bornes de l'intervalle de confiance. On cherche les quantiles d'ordre  $\alpha/2$  et d'ordre  $1 - \alpha/2$  de la distribution. Dans notre cas, la loi approchée de

$$\frac{\hat{b}_j - b_j}{\hat{\sigma}(\hat{b}_j)}$$

est normale centrée réduite. Par conséquent, en employant 0,95 comme probabilité de confiance, on a l'intervalle allant de  $\hat{b}_j - 1,96 \hat{\sigma}(\hat{b}_j)$  à  $\hat{b}_j + 1,96 \hat{\sigma}(\hat{b}_j)$ . L'interprétation est la suivante. Cet intervalle a une probabilité égale à 0,95 de contenir la vraie valeur de  $b_j$ . Au lieu de 1,96, il est préférable d'employer le quantile d'ordre 97,5 % de la distribution de Student à 7 degrés de liberté, soit ici 2,364.



Déterminez l'intervalle de confiance à 95 % pour le coefficient de la variable HAB. Comparez-le avec les sorties de régression rappelées au début de la présente partie d'exercice.



5.4.1 Votre réponse



Faites de même pour la constante et pour le coefficient de SIT.



5.4.2 Votre réponse

## SYNTHÈSE

Dans cette partie, nous avons introduit des mesures d'incertitude sur chacun des coefficients du modèle. Ces mesures d'incertitudes sont appelées des erreurs-types. Elles ont un sens dans le contexte où on considère les données utilisées comme un échantillon prélevé de manière aléatoire dans une population et où ce sont ces coefficients pour la population qui nous intéressent, pas seulement ceux relatifs à l'échantillon.

Pour un coefficient donné, une grande erreur-type traduit une gran-



de incertitude sur le coefficient de régression vrai, c'est-à-dire relatif à la population. Une petite erreur-type traduit une petite incertitude sur le coefficient de régression vrai,

Le test d'une hypothèse de nullité d'un coefficient (toujours dans la population) emploie comme statistique de test le rapport entre le coefficient estimé et son erreur-type. Ce rapport est appelé la *statistique de Student* ou *statistique t*. De même, un intervalle de confiance pour le paramètre vrai se base sur l'estimation du paramètre mais également sur l'erreur-type associée au paramètre estimé.

Nous reviendrons sur une justification empirique de ce qui précède dans l'exercice 2.



**Partie 6** Dans cette partie de l'exercice, les résidus sont inspectés pour vérifier le respect des conditions d'application par le modèle.

### 6.1 LA DÉTERMINATION DES RÉSIDUS

Les résidus sont les différences entre les données et les valeurs ajustées par le modèle (valeurs prédites). Ils sont présentés dans un tableau.

- ⇒ Ouvrez de nouveau le classeur de nom CH07EX01.XLS.
- ⇒ Pour atteindre ce tableau, cliquez sur l'onglet de la feuille Main, pressez F5 et sélectionnez TABLE2. Vous pouvez aussi cliquer sur le lien prévu en haut de la feuille principale Main : «Pour les valeurs ajustées et résidus ».



?

Vérifiez quelques-uns des premiers résidus.

6.1.1 Votre réponse

?

En regardant les résidus sous cette forme, peut-on conclure quelque chose ?

6.1.2 Votre réponse



### 6.2 ANALYSE DES VALEURS AJUSTEES ET DES RESIDUS

- ⇒ Sans fermer le classeur de nom CH07EX01.XLS, ouvrez le classeur CH07EX01Regression.XLS.

?

Comparez quelques-uns des nombres de la colonne  $y_i^*$  ci-dessus avec les résultats de la colonne Predicted Y du nouveau classeur. Comparez quelques-uns des nombres de la colonne  $e_i$  avec les résultats de la colonne Residuals du nouveau classeur. Les résultats sont-ils identiques ?





### 6.2.1 Vos réponses

## 6.3 EXAMEN DES GRAPHIQUES

⇒ Revenez au classeur MP07EX01Regression.XLS que vous avez sauvé. Cliquez une seule fois sur le graphique en haut à gauche (en dessous de la pile) de manière à le faire apparaître. Il peut être utile d'agrandir le graphique en tirant son coin inférieur droit tout en enfonçant le bouton gauche de la souris. Approchez le pointeur de la souris afin de consulter les coordonnées des points.



De quels nombres s'agit-il ? Justifiez le titre du graphique.



### 6.3.1 Votre réponse

⇒ Faites apparaître le troisième graphique. Approchez le pointeur de la souris afin de consulter les coordonnées des points.



De quels nombres s'agit-il ? Justifiez le titre du graphique.



### 6.3.2 Votre réponse

⇒ Faites apparaître le graphique du bas. Approchez le pointeur de la souris afin de consulter les coordonnées des points.



De quels nombres s'agit-il ? Justifiez le titre du graphique.





### 6.3.3 Votre réponse

Il s'agit d'un graphique appelé diagramme quantile normal. La linéarité de la figure correspond à la normalité de la distribution des résidus. Si les points sont autour d'une courbe incurvée et non d'une droite, cela indique que les résidus ne sont pas du tout distribués selon une distribution normale. Si la plupart des points sont sur une droite sauf quelques points aux extrémités, c'est le signe de la présence de données aberrantes.

Nous reviendrons aux résidus dans d'autres exercices.

## SYNTHÈSE

Nous avons examiné les résidus du modèle de régression linéaire multiple de plusieurs façons. Dans les exercices qui suivent, nous verrons que la situation n'est pas toujours aussi bonne que dans les cas traités ici et que ces graphiques peuvent indiquer un manquement plus ou moins grave aux conditions d'application de la méthode de régression linéaire multiple.



**Partie 7** Dans cette partie de l'exercice, un jeu de données légèrement étendu est examiné avec l'utilisation d'une variable de situation à 3 valeurs au lieu de 2, ce qui oblige à introduire des variables binaires dans le modèle.

Il s'agit de la variable SIT qui en plus des valeurs égales à 0 en centre ville et 1 en quartier résidentiel peut aussi être égale à 2 pour les maisons qui se trouvent en périphérie. Les nouvelles données numérotées de 11 à 15 correspondent toutes à  $SIT = 2$ .

### 7.1 PRÉSENTATION DES NOUVELLES DONNÉES

- ⇒ Cliquez sur l'onglet 'Second' du classeur CH07EX01.XLS.
- ⇒ Pour atteindre le tableau des données, pressez F5 et sélectionnez TABLE8. Vous pouvez aussi cliquer sur le lien prévu en haut de la feuille secondaire Second : « Pour les données, taper F5 et sélectionnez TABLE8 ».

**?** Comment pensez-vous qu'ont été créées les variables CENTR, RESID et PERIPH ?



7.1.1 Votre réponse

**?** Pourquoi n'utilise-t-on pas la variable SIT dans un modèle de régression, telle qu'elle est définie ?



7.1.2 Votre réponse

### 7.2 RÉGRESSION AVEC LA VARIABLE SIT

Dans la suite, vous allez devoir employer plusieurs fois l'outil de régression d'Excel, comme dans la partie 1 de l'exercice. Il est recommandé de sélectionner chaque fois l'en-tête de colonne et, par conséquent, de cocher la case Labels. Ne prenez pas la colonne C formée de 1 parmi les variables



explicatives. Par défaut, Excel fournit des régressions avec constante. Pour enlever la constante du modèle, ce que nous demanderons une seule fois, il suffit de cocher la case Constant is zero.

⇒ Effectuez une régression en considérant la PRICE comme variable dépendante, et HAB, SIT comme variables explicatives, en sauvant les résultats dans une nouvelle feuille, nommée 'PriceSit1'.

? Quelles sont les valeurs du coefficient de détermination corrigé et de la statistique de test  $F$  ? Notez-les.

7.2.1 Votre réponse



? Est-ce que les coefficients sont significatifs ? Pourquoi ?

7.2.2 Votre réponse



### 7.3 INTRODUCTION DE NOUVELLES VARIABLES

⇒ Effectuez une régression en considérant la PRICE comme variable dépendante, et HAB, CENTR, RESID et PERIPH comme variables explicatives. (utilisez les colonnes à droite du tableau 8), en sauvant les résultats dans une nouvelle feuille, nommée 'PriceSit2'

#### Remarque

Il ne sert à rien d'insister si vous voyez apparaître un message d'erreur. Si nous faisons cet exercice dans TSE, il y aurait également un message d'erreur.



? Pourquoi y aurait-il un message d'erreur ?

7.3.1 Votre réponse





⇒ Effectuez une régression en considérant la PRICE comme variable dépendante, et HAB, CENTR, RESID et PERIPH comme variables explicatives. (utilisez les données à droite du tableau 8) mais cette fois sans la constante c'est-à-dire en cochant la case Constant is zero, en sauvegardant les résultats dans une nouvelle feuille, nommée 'PriceSit3'



Ecrivez l'équation du modèle en reprenant les statistiques de Student comme après la question 5.2.6.

7.3.2 Votre réponse



### Remarque

Attention : le coefficient de détermination corrigé fourni par Excel est faux dans le cas d'une régression sans constante ; utilisez la formule

$$\bar{R}^2 = R^2 - (1 - R^2) \frac{k-1}{n-k} \text{ pour le calculer.}$$



Quelle sont les valeurs de  $R^2$ , du coefficient de détermination corrigé et de la statistique de test  $F$  ? Notez-les.

7.3.3 Votre réponse



## 7.4 RÉALISATION DE PLUSIEURS RÉGRESSIONS AVEC LES NOUVELLES VARIABLES

Nous allons comparer le modèle précédent avec les trois variables binaires mais sans constante avec trois modèles avec constante et seulement deux des trois variables binaires.



Effectuez la régression avec la variable PRICE comme variable dépendante et HAB, CENTR et RESID comme variables explicatives, en sauvegardant les résultats dans une nouvelle feuille, nommée



'PriceSit4'.

?

Ecrivez l'équation du modèle en reprenant les statistiques de Student comme après la question 5.2.6. Ajoutez la valeur de  $R^2$ .

7.4.1 Votre réponse



⇒

Effectuez la régression avec la variable PRICE comme variable dépendante et HAB, CENTR et PERIPH comme variables explicatives, en sauvant les résultats dans une nouvelle feuille, nommée 'PriceSit5'.

?

Ecrivez l'équation du modèle en reprenant les statistiques de Student comme après la question 5.2.6. Ajoutez la valeur de  $R^2$ .

7.4.2 Votre réponse



⇒

Effectuez la régression avec la variable PRICE comme variable dépendante et HAB, RESID et PERIPH comme variables explicatives, en sauvant les résultats dans une nouvelle feuille, nommée 'PriceSit6'.

?

Ecrivez l'équation du modèle en reprenant les statistiques de Student comme après la question 5.2.6. Ajoutez la valeur de  $R^2$ .

7.4.3 Votre réponse





## 7.5 INTERPRÉTATION DES RÉSULTATS

?

Comparez les résultats du paragraphe précédent. Où voit-on les différences ?

7.5.1 Votre réponse



?

Déterminez au moyen des trois régressions la prévision pour une maison dont HAB vaut 20 et SIT vaut 2. Comparez les prévisions obtenues.

7.5.2 Votre réponse



## SYNTHÈSE

Dans les parties 1 à 6, la situation de la maison était de nature dichotomique c'est-à-dire avec deux valeurs différentes possibles. Ici, nous avons ajouté des données qui correspondent à une troisième situation. Nous avons d'abord montré qu'employer la variable SIT à trois valeurs peut conduire à des résultats étonnants qui dépendent en outre de la codification employée.

Nous avons ensuite décrit une autre approche qui consiste à créer une variable binaire pour chaque valeur de la variable catégorielle et à employer ces différentes variables sauf une. Dans l'exemple, il s'agit de deux variables. Nous avons montré que le choix des deux variables parmi les trois variables possibles n'a pas d'effet sur les valeurs ajustées par le modèle même si cela présente un effet sur les résultats estimés.

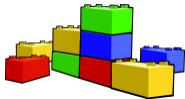




## Exercice avancé

(Pour les utilisateurs de la version avancée du cours)

### Préalable



Le chapitre 7 du cours de base et du cours avancé doit avoir été suivi jusqu'à la page 48, pour la partie A, jusqu'à la page 49, pour la partie B, jusqu'à la page 68, pour la partie C, jusqu'à la page 80, pour la partie D, jusqu'à la page 50, pour la partie E (facultative).

### Objectif



Le but est ici de discuter les aspects de calcul associés à la méthode des moindres carrés pour la régression linéaire multiple ainsi que certains aspects théoriques (à titre facultatif pour ces derniers).

### Données



Les mêmes données que dans l'exercice de base.



### Structure de l'exercice

L'exercice avancé comporte cinq parties :

- Dans la partie A, le but de l'exercice est de décrire sur l'exemple, mais de manière détaillée, le calcul des coefficients du modèle de régression linéaire multiple estimés par la méthode des moindres carrés. On emploie le calcul matriciel à l'aide d'Excel.
- Dans la partie B, le but de l'exercice est de discuter les aspects numériques de calcul et les problèmes posés par la colinéarité et la quasi-colinéarité.
- Dans la partie C, le but de l'exercice est compléter la partie A en décrivant la manière d'obtenir la variance résiduelle et le coefficient de détermination comme sous-produits du calcul des coefficients estimés.
- Dans la partie D, le but de l'exercice est de compléter la partie A en décrivant la manière d'obtenir les erreurs-types et les statistiques pour les tests d'hypothèse ainsi que la détermination des intervalles de confiance sur les paramètres.
- Dans la partie E, le but de l'exercice *facultatif* est de fournir les justifications *théoriques* de la méthode des moindres carrés et des formules utilisées dans les parties A à D de l'exercice avancé. Ces justifications théoriques reposent non seulement sur la minimisation d'une fonction de plusieurs variables par considération des dérivées de cette fonction, comme dans la régression linéaire simple, mais également sur des éléments de calcul matriciel.



**Partie A** Dans cette partie, le but de l'exercice est de décrire sur l'exemple, mais de manière détaillée, le calcul des coefficients du modèle de régression linéaire multiple estimés par la méthode des moindres carrés.

Les justifications *théoriques* de la méthode des moindres carrés et des formules utilisées sont données dans la partie E, à titre *facultatif*. Ici, on se contente d'observer les formules dans Excel et de comparer les résultats avec ceux des parties 1 et 3 du cours de base. On fournit à cette fin quelques éléments de calcul matriciel.

L'exercice est facultatif dans la mesure où nous n'aurons jamais besoin d'appliquer ces formules qui sont directement appliquée dans les logiciels de bureautique ou les logiciels statistiques.

### Remarques



1. Le point important est d'être conscient de l'existence des formules qui fournissent une solution unique et (presque toujours) exacte.
2. Pour éclairer quelque peu la restriction exprimée à la remarque précédente, l'approche suivie ici qui correspond à l'application des formules théoriques n'est pas la mieux adaptée au calcul numérique. Cet aspect sera mieux argumenté dans l'exercice 6.

### A.a PRÉSENTATION MATRICIELLE DES DONNÉES : LA MATRICE X

On considère l'exemple des parties 1 et 2, avec 10 données.



Dans la feuille Main, pressez F5 et sélectionnez X.



Expliquez comment la matrice X est construite.



A.a.1 Votre réponse



Pourquoi la première colonne est-elle formée par des 1 ?



A.a.2 Votre réponse



**A.b PREMIÈRE ÉTAPE : TRANSPOSITION DE X**

Pressez F5 et sélectionnez XT.



Comment la matrice X est-elle transformée en XT (le transposé ou la transposée de X) ?



*A.b.1 Votre réponse*

La formule utilisée dans la plage H30 : Q32 emploie la fonction TRANSPOSE qui est une fonction matricielle, *array function*. On voit la formule {=TRANSPOSE ( S19 : U28 ) }.

**Remarque**

ATTENTION. Quand on tape une fonction matricielle, il ne faut surtout pas taper les accolades. Il ne faut pas non plus employer la touche Enter pour valider la formule mais une combinaison de trois touches Ctrl Shift Enter. Sans cela, la formule n'est pas prise en compte de manière adéquate. **Cette remarque s'applique à toutes les formules matricielles.**

**A.c SECONDE ÉTAPE : PRODUIT X TRANSPOSÉ PAR X**

Pressez F5 et sélectionnez XTX.

Avant de passer à la question suivante, donnons (ou rappelons) la définition d'un produit scalaire. Supposons qu'on ait ce qu'on appelle un vecteur ligne ou une matrice ligne à trois composantes, par exemple (1 3 2) et un vecteur ou une matrice colonne à trois composantes, par

exemple  $\begin{pmatrix} 4 \\ 6 \\ 5 \end{pmatrix}$ . On appelle le produit scalaire de la ligne par la colonne la

somme des produits des éléments correspondants, c'est-à-dire le premier de la ligne avec le premier de la colonne, le deuxième de la ligne avec le deuxième de la colonne et le troisième de la ligne avec le troisième de la colonne, par conséquent :  $1 \times 4 + 3 \times 6 + 2 \times 5 = 4 + 18 + 10 = 32$ . Notez qu'on doit toujours avoir une ligne à gauche et une colonne à droite et qu'elles doivent comporter le même nombre d'éléments.





?

Multipliez la ligne 1 de  $XT$  par la colonne 1 de  $X$  en faisant le produit scalaire. Comparez le résultat avec la valeur en ligne 1 et colonne 1 de la matrice  $XTX$ .

*A.c.1 Votre réponse*



?

Multipliez la ligne 2 de  $XT$  par la colonne 3 de  $X$  en faisant le produit scalaire. Comparez le résultat avec la valeur en ligne 2 et colonne 3 de la matrice  $XTX$ .

*A.c.2 Votre réponse*



?

Vérifiez les autres éléments de la matrice  $XTX$ . Sont-ils obtenus de la même manière ?

*A.c.3 Votre réponse*



### Remarque

Le produit d'une matrice  $A$  à  $m$  lignes et  $n$  colonnes par une matrice  $B$  à  $n$  lignes et  $k$  colonnes donne la matrice  $C$  à  $m$  lignes et  $k$  colonnes telle que la valeur en  $i$ ème ligne et  $j$ ème colonne est la produit (au sens du produit scalaire défini plus haut) de la  $i$ ème ligne de la matrice  $A$  par la  $j$ ème colonne de la matrice  $B$ . C'est ce qu'on appelle le *produit matriciel*. Notons que le produit matriciel n'a de sens que si le nombre de colonnes de la matrice  $A$  est égal au nombre de lignes de la matrice  $B$ , ici le nombre  $n$ .

Le calcul est effectué dans la plage S30 : U32 à l'aide de la fonction de produit matriciel MMULT,  $\{=MMULT(H30:Q32;S19:U28)\}$ . Les deux arguments sont la plage contenant  $XT$ , d'une part, la plage contenant  $X$ , d'autre part.



**Remarque**

Rappelons qu'il ne faut pas taper les accolades pour une fonction matricielle mais qu'il faut la valider avec la combinaison de touches Ctrl Shift Enter, au lieu de Enter.

**A.d TROISIÈME ÉTAPE : INVERSION**

Pressez F5 et sélectionnez XTXM1.

La matrice XTX étant  $\begin{pmatrix} 10 & 220 & 3 \\ 220 & 5006 & 69 \\ 3 & 69 & 3 \end{pmatrix}$ , on veut calculer l'inverse de cette

matrice. L'inverse d'une matrice  $A$  est une matrice  $B$ , souvent notée  $A^{-1}$ , telle que le produit matriciel de  $A$  par  $B$  (ou de  $B$  par  $A$ ) soit la matrice identité, souvent notée  $I$ , qui comporte des zéros partout sauf sur la diagonale principale (ou descendante) qui contient des 1. Donc ici

$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ . Ceci s'effectue dans Excel à l'aide d'une fonction matricielle

MINVERSE. Plus précisément, dans la plage S35 :U37, on voit la formule  $\{=MINVERSE(S30:U32)\}$ .

**Remarques**

1. Rappelons qu'il ne faut pas taper les accolades pour une fonction matricielle mais qu'il faut la valider avec la combinaison de touches Ctrl Shift Enter, au lieu de Enter.

2. L'inversion d'une matrice peut aussi se faire de bien des manières. Ne parlons pas de l'emploi des déterminants, tout à fait inefficace. La méthode d'élimination de Gauss peut être employée. Une autre manière consiste à utiliser les formules d'inversion suivantes :

$$\text{si } M = \begin{pmatrix} M_{11} & | & M_{12} \\ --- & -|- & --- \\ M_{21} & | & M_{22} \end{pmatrix},$$

alors

$$M^{-1} = \begin{pmatrix} M^{11} & | & M^{12} \\ --- & -|- & --- \\ M^{21} & | & M^{22} \end{pmatrix},$$

avec, successivement,



$$M^{22} = (M_{22} - M_{21}M_{11}^{-1}M_{12})^{-1}$$

$$M^{11} = M_{11}^{-1} + M_{11}^{-1}M_{12}M^{22}M_{21}M_{11}^{-1}$$

$$M^{12} = -M_{11}^{-1}M_{12}M^{22}$$

$$M^{21} = -M^{22}M_{21}M_{11}^{-1}$$

On peut vérifier le calcul de la matrice inverse en demandant le produit de la matrice  $XTX$  par son inverse  $XTXM1$ . C'est ce qui a été fait dans la plage S39 :U41.

?

Vérifiez la formule employée. Que pensez-vous du résultat obtenu ?



*A.d.1 Votre réponse*

#### A.e QUATRIÈME ÉTAPE : OBTENTION DES ESTIMATIONS



Pour voir le vecteur  $X'Y$  pressez F5 et sélectionnez XTY.

?

Vérifiez la formule employée ainsi que le résultat relatif au premier et au troisième éléments. Comment pourrait-on interpréter les trois éléments de  $X'Y$  ?



*A.e.1 Votre réponse*



Pour voir le vecteur  $\hat{b}$  qu'on veut estimer, descendez un peu.

?

Il se trouve sous le titre  $INV(X'X).X'Y$ . Pourquoi à votre avis ? Vérifiez la formule utilisée et le premier élément.



*A.e.2 Votre réponse*





?

Comparez les trois estimations avec les valeurs trouvées dans l'annexe de la partie 1, par exemple. Est-ce identique ?

*A.e.3 Votre réponse*

### Remarque



En régression linéaire simple, nous avons vu que le coefficient de régression s'obtient à partir de la variance de la variable explicative

$$s_x^2 = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})^2]$$

et de la covariance entre les  $x_i$  et les  $y_i$  est donnée par l'expression

$$s_{xy} = \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]$$

en effectuant le rapport :

$$b_1 = \frac{\text{cov}(x, y)}{\text{var}(x)} = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

Cette formule ne se généralise pas directement à la régression multiple.

Néanmoins, on peut encore dire que le plan (ou l'hyperplan si  $k > 2$ ) des moindres carrés passe par le *centre de gravité* des données. Dans l'exemple, les moyennes des trois variables sont PRICE :  $370/10 = 37$  ; HAB :  $220/10 = 22$  ; SIT :  $3/10 = 0,3$ . On a bien  $18,28 + 0,725 \times 22 + 9,20 \times 0,3 = 37$ .

## SYNTHÈSE

Nous avons montré comment effectuer les calculs détaillés pour l'estimation par la méthode des moindres carrés pour la régression linéaire multiple.



**Partie B** Dans cette partie, le but de l'exercice est de discuter les aspects numériques de calcul et les problèmes posés par la colinéarité et la quasi-colinéarité.

À cette fin, nous partons des formules employées dans la partie A et nous allons simplement modifier les données. L'exercice peut être effectué de deux manières : soit en employant l'outil de régression comme dans la partie 1, mais il faut s'attendre à voir peu de résultats, soit en employant le calcul matriciel comme dans la partie A. Dans ce dernier cas, il faut copier les valeurs des colonnes dans les colonnes appropriées de la plage X et observer la mise à jour des calculs. Observez la matrice inverse et ce qui doit être la matrice identité (avec des zéros sauf des 1 sur la diagonale principale). Une macro-instruction de nom CH07EX01Initial permet de remettre les données d'origine dans les plages X et Y.

### B.a RÉGRESSION AVEC DEUX FOIS LA MÊME VARIABLE

⇒ Dans la feuille Main descendez jusqu'à la ligne 72. Dans ce tableau la variable HAB est recopiée sous le nom de HAB1.

? Effectuez une régression de PRICE sur les variables HAB, HAB1, SIT. Trouvez-vous un résultat ? Pourquoi ?



*B.a.1 Votre opinion*

### B.b RÉGRESSION AVEC UNE VARIABLE SOMME DE DEUX AUTRES

⇒ Dans la feuille main descendez jusqu'à la ligne 86. Dans ce tableaux on a ajoute la variable SOM qui est la somme de variables HAB et SIT.

? Effectuez une régression de PRICE sur les variables HAB, HAB1, SIT. Trouvez-vous un résultat ? Pourquoi ?





*B.b.1 Votre opinion*

### **B.c RÉGRESSION AVEC COLINÉARITE**



Pourquoi la régression ne fonctionne-t-elle pas dans les paragraphes précédents ?



*B.c.1 Votre réponse*

### **B.d RÉGRESSION AVEC PERTURBATION DE VARIABLES COLINÉAIRES**



Dans la feuille main descendez jusqu'à la ligne 101. Dans ce tableau la variable HAB est recopiée avec quelque légères perturbations sous le nom HAB2.



Effectuez une régression de PRICE sur les variables HAB, HAB2, SIT. Est-ce que les coefficients de HAB et HAB2 sont significatifs dans ce modèle ? Pourquoi ?



*B.d.1 Votre opinion*

### **B.e RÉFLEXIONS SUR L'INDÉTERMINATION DES COEFFICIENTS**

Supposons qu'on veuille expliquer une variable  $y$  par une régression linéaire, avec les variables explicatives deux fois  $x$ .

Soient  $x_1, x_2, x_3$  les observations de  $x$  qui sont deux fois dans le modèle en plus de la constante et  $y_1, y_2, y_3$  les observations de  $y$ .



**?**

Comment s'écrivent la matrice  $X$  et la matrice transposée de  $X$  ?

*B.e.1 Votre réponse*

$$X = \begin{pmatrix} & \\ & \end{pmatrix} \quad X' = \begin{pmatrix} & \\ & \end{pmatrix}$$

**?**

Calculez la matrice  $XX'$ . Est-elle inversible ? Pourquoi ?

*B.e.2 Votre réponse*

$$XX' = \begin{pmatrix} & \\ & \end{pmatrix}$$

## SYNTHÈSE

Nous avons montré une condition indispensable pour l'utilisation de la régression linéaire multiple : les variables explicatives doivent être sans relation linéaire entre elles. En cas d'existence d'une relation linéaire, les coefficients de régression deviennent indéterminés et ne peuvent pas être calculés de manière unique. Nous avons montré le caractère intuitif de cette règle. Nous avons également traité du cas de la quasi-colinéarité, le cas où il n'y a pas de relation linéaire exacte entre les variables explicatives mais bien une relation statistique intense. Cette situation se révèle pire parce qu'elle n'empêche pas l'obtention de résultats numériques qui peuvent paraître sensés à première vue mais qui sont souvent dénués de toute signification.



**Partie C** Dans cette partie, le but de l'exercice est de compléter la partie A en décrivant la manière d'obtenir la variance résiduelle et le coefficient de détermination comme sous-produits du calcul des coefficients estimés.

### C.a OBTENTION DE LA VARIANCE RÉSIDUELLE

Considérons l'exemple des maisons avec 10 données, voir partie 2. Nous avons déjà vérifié que les résidus de la plage F47 : F56 sont corrects dans le tableau 2.

⇒ Pour atteindre ce tableau, cliquez sur l'onglet de la feuille Main, pressez F5 et sélectionnez TABLE2. Vous pouvez aussi cliquer sur le lien prévu en haut de la feuille principale Main : «Pour les valeurs ajustées et résidus ».

Dans la cellule E30, on a calculé MSE (qui est aussi la variance résiduelle *sans* correction pour le nombre de degrés de liberté, c'est-à-dire avec  $n$  au dénominateur, et non  $n - 1$  pour les variances corrigées comme habituellement). On emploie à cette fin la fonction d'Excel VARP (pour variance population).

Dans la cellule E31, on a calculé la variance résiduelle *avec* correction pour le nombre de degrés de liberté, c'est-à-dire avec  $n - k$  au dénominateur (et non  $n - 1$  comme habituellement). Nous la noterons  $\hat{\sigma}^2$ . On a également calculé sa racine carrée  $\hat{\sigma}$ , l'écart-type résiduel, dans la cellule F31.

? Vérifiez les formules. Retrouvez la variance résiduelle dans la feuille 'PriceHabSit' ou en annexe de la partie 1. Comparez également le résultat de l'écart-type résiduel.



C.a.1 Votre réponse

### C.b OBTENTION DU COEFFICIENT DE DÉTERMINATION

On a calculé la variance de la variable dépendante  $s_y^2$ , également sans correction dans la cellule B30, puis avec correction dans la cellule B31.



?

On ne trouve pas la variance des prix dans la feuille 'Price-HabSit' ou en annexe de la partie 1 mais bien la somme des carrés correspondante, le numérateur. Où se trouve-t-elle ?

*C.b.1 Votre réponse*

?

Sachant que la  $R^2 = 1 - \text{MSE}/s_y^2$  est calculé dans la cellule G30, vérifiez sa valeur et comparez le résultat avec les sorties obtenues dans la feuille 'PriceHabSit' ou en annexe de la partie 1.

*C.b.2 Votre réponse*



Nous avons introduit le coefficient de détermination corrigé  $\bar{R}^2$  dans la partie 4, plus précisément au paragraphe 4.2. Nous l'avons défini par la formule suivante :

$$\bar{R}^2 = R^2 - (1 - R^2) \frac{k-1}{n-k}.$$

Nous avons dit que son introduction se justifie en considérant la régression dans le contexte d'une population plutôt que de celui d'un échantillon issu de celle-ci. Au lieu des variances échantillon, comportant le nombre d'observations  $n$  au dénominateur, on est alors tenté d'employer des estimations non biaisées des variances dans la population. Ces estimations sont donc corrigées pour le biais en tenant compte des nombres de degrés de liberté adéquats. Pour la variance d'y, il suffit d'employer  $n - 1$  au lieu de  $n$ . Pour la variance résiduelle, obtenue en estimant les  $k$  coefficients de régression, le nombre de degré de liberté est  $n - k$  au lieu de  $n$ .  $\bar{R}^2$  est la solution, exprimée en termes de  $R^2$  de l'emploi d'une équation similaire à celle trouvée pour  $R^2$ , en remplaçant les deux variances par des estimations non biaisées :

$$\hat{\sigma}^2 = \frac{\text{MSE}}{n-k} = \frac{n}{n-1} s_y^2 (1 - \bar{R}^2)$$

C'est ce qu'on peut vérifier dans la cellule G31.



**?**

Expliquez l'interprétation qu'on peut donner à  $\bar{R}^2$ . Comparez le résultat avec les sorties obtenues dans la feuille 'PriceHabSit' ou de l'annexe de la partie 1.

*C.b.3 Votre réponse*

## SYNTHÈSE

Nous avons montré comment obtenir MSE, la variance résiduelle et le coefficient de détermination.



**Partie D** Dans cette partie, le but de l'exercice est de compléter la partie A en décrivant la manière d'obtenir les erreurs-types et les statistiques pour les tests d'hypothèse ainsi que la détermination des intervalles de confiance sur les paramètres.

Pour cela considérons le modèle  $y = b_0 + b_1x_1 + \dots + b_kx_k + e$ .

### D.a OBTENTION DES ERREURS-TYPES

L'erreur-type pour l'estimation du paramètre  $b_j$  s'obtient comme racine carrée de la formule  $\hat{\sigma}^2(\hat{b}_j) = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})_{jj}^{-1}$ , où  $(\mathbf{X}'\mathbf{X})_{jj}^{-1}$  représente l'élément  $(j, j)$  de la matrice inverse de  $\mathbf{X}'\mathbf{X}$  et  $\hat{\sigma}^2$  est la variance résiduelle.

⇒ Dans la feuille Main appuyez sur F5 et cliquez sur XTXM1 pour avoir l'inverse de la matrice  $\mathbf{X}'\mathbf{X}$ .

La valeur de  $\hat{\sigma}^2$  est reprise dans la cellule S45.



Vérifiez que les formules employées pour calculer les trois erreurs-types, dans les cellules S42, T43 et U44 sont bien conformes à ce qui précède.



*D.a.1 Votre réponse*



Comparez le résultat avec les sorties obtenues dans la feuille 'PriceHabSit' ou en annexe de la partie 1.



*D.a.2 Votre réponse*

### D.b DÉTERMINATION DES STATISTIQUES DE TEST

La statistique  $t$  s'obtient par la formule  $t = \hat{b}_j / \hat{\sigma}(\hat{b}_j)$ .





Dans la feuille Main appuyez sur F5 et cliquez sur BETA. Descendez un peu.

Les statistiques de Student ont été calculées dans la plage W42 : W44.



Vérifiez les formules utilisées et les résultats des statistiques  $t$ . Comparez les résultats avec les sorties obtenues dans la feuille 'PriceHabSit' ou l'annexe de la partie 1.



*D.b.1 Votre réponse*

### **D.c CONSTRUCTION DES INTERVALLES DE CONFIANCE POUR LES PARAMÈTRES**

L'intervalle de confiance à 95% s'obtient par la formule  $\hat{b}_j \pm t\hat{\sigma}(\hat{b}_j)$  où  $t$  est la valeur du quantile d'ordre 97,5% d'une distribution normale centrée réduite.



Calculez l'intervalle de confiance à 95% pour le coefficient de la variable SIT. Comparez le résultat avec les sorties obtenues dans la feuille 'PriceHabsit' ou de l'annexe de la partie 1.



*D.c.1 Votre réponse*

### **D.d OBTENTION DES STATISTIQUES DE TEST POUR LE MODÈLE**

La statistique de test pour le modèle global s'obtient par la formule 
$$\frac{R^2}{\frac{k-1}{1-R^2} \cdot \frac{1}{n-k}}$$





Vérifiez la valeur de la statistique  $F$  et comparez le résultat avec les sorties obtenues dans la feuille 'PriceHabSit' ou de l'annexe de la partie 1.

*D.d.1 Votre réponse*

### D.e OBTENTION DES STATISTIQUES DE TEST POUR UN SOUS-MODÈLE

On peut vérifier de même (et les développements sont trop complexes pour être traités ici) que la statistique de test pour tester l'hypothèse nulle  $b_1 =$

$b_2 = \dots = b_q = 0$ , s'obtient par la formule  $\frac{\frac{R_c^2 - R_r^2}{q}}{\frac{1 - R_c^2}{n - k}}$  où  $R_c^2$  est le coefficient de

détermination du modèle complet, avec toutes les variables :

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_{k-1}x_{k-1} + e,$$

et  $R_r^2$  est le coefficient de détermination du modèle restreint, sans les variables  $x_1, x_2, \dots, x_q$ , c'est-à-dire sous l'hypothèse  $b_1 = b_2 = \dots = b_q = 0$  :

$$y = b_0 + b_{q+1}x_{q+1} + b_{q+2}x_{q+2} + \dots + b_{k-1}x_{k-1} + e.$$

Il faut donc effectuer deux régressions. Considérons par exemple le test que le coefficient de SIT est nul. Le modèle de régression sans la variable SIT a été déterminé dans la partie 1 de l'exercice. Il s'obtient en cliquant sur l'onglet PriceHab.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.54638219
R Square	0.2985335
Adjusted R Squa	0.21085019
Standard Error	6.22543365
Observations	10

ANOVA

	df	SS	MS	F	Significance F
Regression	1	131.9518072	131.952	3.40468	0.102230374
Residual	8	310.0481928	38.756		
Total	9	442			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%
Intercept	17.3855422	10.81087852	1.60815	0.14647	-7.544404523
HAB	0.89156627	0.48318736	1.84518	0.10223	-0.222666505



On associe à la différence  $R_c^2 - R_r^2$  un nombre de degrés de liberté égal à  $k - 1 - q + 1 = k - q$ , puisqu'il y a  $k - q$  coefficients y compris la constante de régression. Il est certain que  $R_r^2$  est inférieur ou égal à  $R_c^2$ . On associe à la différence  $1 - R_r^2$  un nombre de degrés de liberté égal à  $k$ , puisqu'il y a  $k$  coefficients estimés.

?

Calculez la statistique  $F$  sur base des coefficients de détermination des deux régressions ? Quels sont les nombres de degrés de liberté associés ?.



*D.e.1 Votre réponse*

?

Comparez avec la statistique correspondante au coefficient de SIT dans la feuille 'PriceHabSit' (ou le tableau de l'annexe de la partie 1) ou plutôt à son carré.



*D.e.2 Votre réponse*

### Remarque



Le carré d'une variable aléatoire ayant une distribution de Student à  $n - k$  degrés de liberté a une distribution de Fisher à 1 et  $n - k$  degrés de liberté.

## SYNTHÈSE

Nous avons montré comment calculer les statistiques de Student comme sous-produit des calculs développés dans la partie A. C'est en fait, l'obtention des erreurs-types qui s'en déduisent directement, avec également la possibilité de déterminer les intervalles de confiance pour les paramètres.

Le test sur le modèle complet ou sur un sous-modèle nécessite l'estimation des paramètres de deux modèles de régression. Nous décrivons la méthode générale, ce qui peut s'avérer d'autant plus utile que très peu de logiciels statistiques (et aucun logiciel de bureautique) ne permet ceci de manière simple.



**Partie E** *Facultatif*

Dans cette partie, le but de l'exercice *facultatif* est de fournir les justifications *théoriques* de la méthode des moindres carrés et des formules utilisées dans les parties A à D de l'exercice avancé. Ces justifications théoriques reposent non seulement sur la minimisation d'une fonction de plusieurs variables par considération des dérivées de cette fonction, comme dans la régression linéaire simple, mais également sur des éléments de calcul matriciel. Afin de garder des formules simples, nous traitons le cas de trois paramètres estimés mais la méthode se généralise directement à plus de paramètres.

**E.a OBJECTIF**

On considère un modèle de régression à deux variables explicatives en plus de la constante suivant :  $y = b_0 + b_1x_1 + b_2x_2 + e$ .

**E.b OBTENTION DES ÉQUATIONS DITES NORMALES**

On peut écrire le modèle sous la forme  $Y = X'b + e$ ,

$$\text{avec } b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \text{ et } e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix}.$$

Supposons qu'on dispose de  $n$  données relatives aux 3 variables,  $y$  et les  $x_1$ ,  $x_2$ , qu'on représente dans un vecteur  $y$  et dans une matrice  $X$  :

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1n} & x_{2n} \\ 1 & x_{1n} & x_{2n} \\ \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$$

Supposons qu'on ait réussi à estimer les coefficients de régression. On peut alors calculer les valeurs ajustées  $y^*$ ,  $i = 1, \dots, n$ , et les résidus  $e_i$  par les formules :

$$y^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \dots \\ y_n^* \end{pmatrix} = \begin{pmatrix} b_0 + b_1x_{21} + b_2x_{21} \\ b_0 + b_1x_{12} + b_2x_{22} \\ \dots \\ b_0 + b_1x_{1n} + b_2x_{2n} \end{pmatrix}, \quad e = \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_n \end{pmatrix} = \begin{pmatrix} y_1 - y_1^* \\ y_2 - y_2^* \\ \dots \\ y_n - y_n^* \end{pmatrix}.$$

Ces formules peuvent être réécrites à l'aide du calcul matriciel. En effet, en notant  $X_i$  la  $i$ ème ligne de  $X$ , on a :



$$y_i^* = b_0 + b_1 x_{1i} + b_2 x_{2i} = \begin{pmatrix} 1 & x_{1i} & x_{2i} \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \mathbf{X}_i \mathbf{b},$$

d'où

$$\mathbf{y}^* = \mathbf{X}\mathbf{b}$$

$$\mathbf{e} = \mathbf{y} - \mathbf{y}^* = \mathbf{y} - \mathbf{X}\mathbf{b}.$$

L'estimation des paramètres peut se baser sur le principe des moindres carrés ("*least squares*"), qui consiste à minimiser

$$Q(b_0, b_1, b_2) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left[ y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i}) \right]^2.$$

En égalant à zéro les dérivées partielles de  $Q$  par rapport à  $b_0, b_1, b_2$ , on trouve un système d'équations linéaires en  $b_0, b_1, b_2$ , de la forme :

$$\begin{aligned} nb_0 + b_1 \left( \sum_i x_{1i} \right) + b_2 \left( \sum_i x_{2i} \right) &= \sum_i y_i \\ b_0 \left( \sum_i x_{1i} \right) + b_1 \left( \sum_i x_{1i}^2 \right) + b_2 \left( \sum_i x_{1i} x_{2i} \right) &= \sum_i x_{2i} y_i \\ b_0 \left( \sum_i x_{2i} \right) + b_1 \left( \sum_i x_{2i} x_{1i} \right) + b_2 \left( \sum_i x_{2i}^2 \right) &= \sum_i x_{2i} y_i. \end{aligned}$$

On dit que l'*hyperplan de régression* au sens des moindres carrés passe par le *centre de gravité*, point de coordonnées  $(\bar{x}_1, \bar{x}_2, \dots, \bar{y})$  de l'espace à  $k$  dimensions  $(x_1, x_2, \dots, y)$ . On peut vérifier que les autres coefficients peuvent ensuite s'obtenir en résolvant un système de  $k - 1$  équations où l'on remplace  $y_i$  par  $y_i - \bar{y}$ ,  $x_{1i}$  par  $x_{1i} - \bar{x}_1$ , etc., pour  $i = 1, \dots, n$ .

Le système d'équations linéaire s'appelle *système des équations normales*.

### E.c ECRITURE DES ÉQUATIONS NORMALES SOUS FORME MATRICIELLE

Le système peut se mettre sous forme matricielle :  $\mathbf{M}\mathbf{b} = \mathbf{m}$ , où

$$\mathbf{M} = \begin{pmatrix} n & \sum_i x_{1i} & \sum_i x_{2i} x_{1i} \\ \sum_i x_{1i} & \sum_i x_{1i}^2 & \sum_i x_{1i} x_{2i} \\ \sum_i x_{2i} & \sum_i x_{1i} x_{2i} & \sum_i x_{2i}^2 \end{pmatrix}, \quad \mathbf{m} = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{1i} y_i \\ \sum_i x_{2i} y_i \end{pmatrix}.$$

Notons que  $\mathbf{m}$  peut se mettre sous la forme d'un produit matriciel

$$\mathbf{m} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \mathbf{X}' \mathbf{y},$$



De même,  $M$  peut s'écrire  $M = \mathbf{X}'\mathbf{X}$ .

### E.d PRÉSENTATION EN EMPLOYANT L'INVERSION MATRICIELLE

Les coefficients de régression s'obtiennent donc comme solution du système  $(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$

Si ce système possède une solution unique, la matrice  $\mathbf{M}$  carrée  $3 \times 3$  est inversible, c'est-à-dire qu'il existe une matrice  $3 \times 3$  notée  $\mathbf{M}^{-1}$  telle que  $\mathbf{M}\mathbf{M}^{-1} = \mathbf{I}$ , la matrice unité  $3 \times 3$ , qui comporte partout des zéros sauf sur la diagonale principale qui est remplie de 1. On note alors la solution fournie par la méthode des moindres carrés  $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$ .

### E.e CAS PARTICULIER DE LA RÉGRESSION LINÉAIRE SIMPLE

Le modèle étudié est le suivant :  $y = b_0 + b_1x + e$ . On a

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}.$$

L'estimateur recherché est  $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$

D'après ce qui précède, on a  $\mathbf{X}' = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{pmatrix}$  et  $\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$ , donc

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{pmatrix}, \text{ et } (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

On a aussi  $\mathbf{X}'\mathbf{y} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{pmatrix}$ , donc

$$\hat{\mathbf{b}} = \frac{1}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n y_i x_i \end{pmatrix}$$



$$\begin{aligned}
&= \frac{1/n^2}{s_x^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i \\ n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i \end{pmatrix} = \frac{1}{s_x^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 \bar{y} - \bar{x} \frac{1}{n} \sum_{i=1}^n x_i y_i \\ \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y} \end{pmatrix} \\
&= \frac{1}{s_x^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n x_i^2 \bar{y} - \bar{x}^2 \bar{y} - \bar{x} \frac{1}{n} \sum_{i=1}^n x_i y_i + \bar{x}^2 \bar{y} \\ s_{xy}^2 \end{pmatrix} = \frac{1}{s_x^2} \begin{pmatrix} s_x^2 \bar{y} - s_{xy}^2 \bar{x} \\ s_{xy}^2 \end{pmatrix} = \begin{pmatrix} \bar{y} - \frac{s_{xy}^2}{s_x^2} \bar{x} \\ \frac{s_{xy}^2}{s_x^2} \end{pmatrix}
\end{aligned}$$

ce qui est le résultat du chapitre 2 pour le modèle linéaire simple :

$$\hat{b}_1 = \frac{s_{xy}^2}{s_x^2} \text{ et } \hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}.$$

L'erreur-type associée à  $b_1$  vaut  $\hat{\sigma}$  fois la racine carrée de l'élément (2,2) de la matrice  $(X'X)^{-1}$ . Cet élément vaut, après simplification de  $n$ , de  $1/(n s_x^2)$ , donc l'erreur-type est égale à  $\hat{\sigma}/(\sqrt{n} s_x)$ .

## SYNTHÈSE

Ces justifications théoriques reposent sur la minimisation d'une fonction de plusieurs variables par considération des dérivées de cette fonction. Non seulement elle fournit une justification des formules employées explicitement ou au travers de logiciels mais aussi elle montre le caractère unique de la solution obtenue à la seule condition d'inversibilité d'une matrice déduite des valeurs des variables explicatives.

[Retour au chapitre 7](#)