

A photograph of the HEC Montréal building at night. The building is a modern, multi-story structure with a glass facade and a blue sky in the background. The text "HEC MONTRÉAL" is visible in the top left corner of the image.

HEC MONTRÉAL

# Modèles d'aide à la décision 4-600-04

## Séance 5 Régression linéaire

# Plan

- 5.1 Introduction
- 5.2 Régression linéaire simple
  - Modèle, interprétation, hypothèses, vocabulaire
  - Tests et signification
  - Inférence, extrapolation
- 5.3 Régression linéaire multiple
- 5.4 Choix des variables explicatives
- 5.5 Variables de catégorie
- 5.6 Relations non-linéaires

## 5.1 Introduction

- L'analyse de régression est utilisée pour estimer une fonction  $f(\cdot)$  qui décrit la relation entre une variable continue dépendante et une ou plusieurs variables indépendantes

$$Y = f(X_1, X_2, X_3, \dots, X_n) + \varepsilon$$

**Note :**

- \*  $f(\cdot)$  décrit la variation systématique entre la variable dépendante et les variables indépendantes (expliquée par la régression).
- \*  $\varepsilon$  représente la variation aléatoire non-expliquée par la régression.

## 5.1 Type de relation statistique

**Y**

Courbe de  
régression

Distributions de probabilité  
pour Y à différents niveaux  
de X

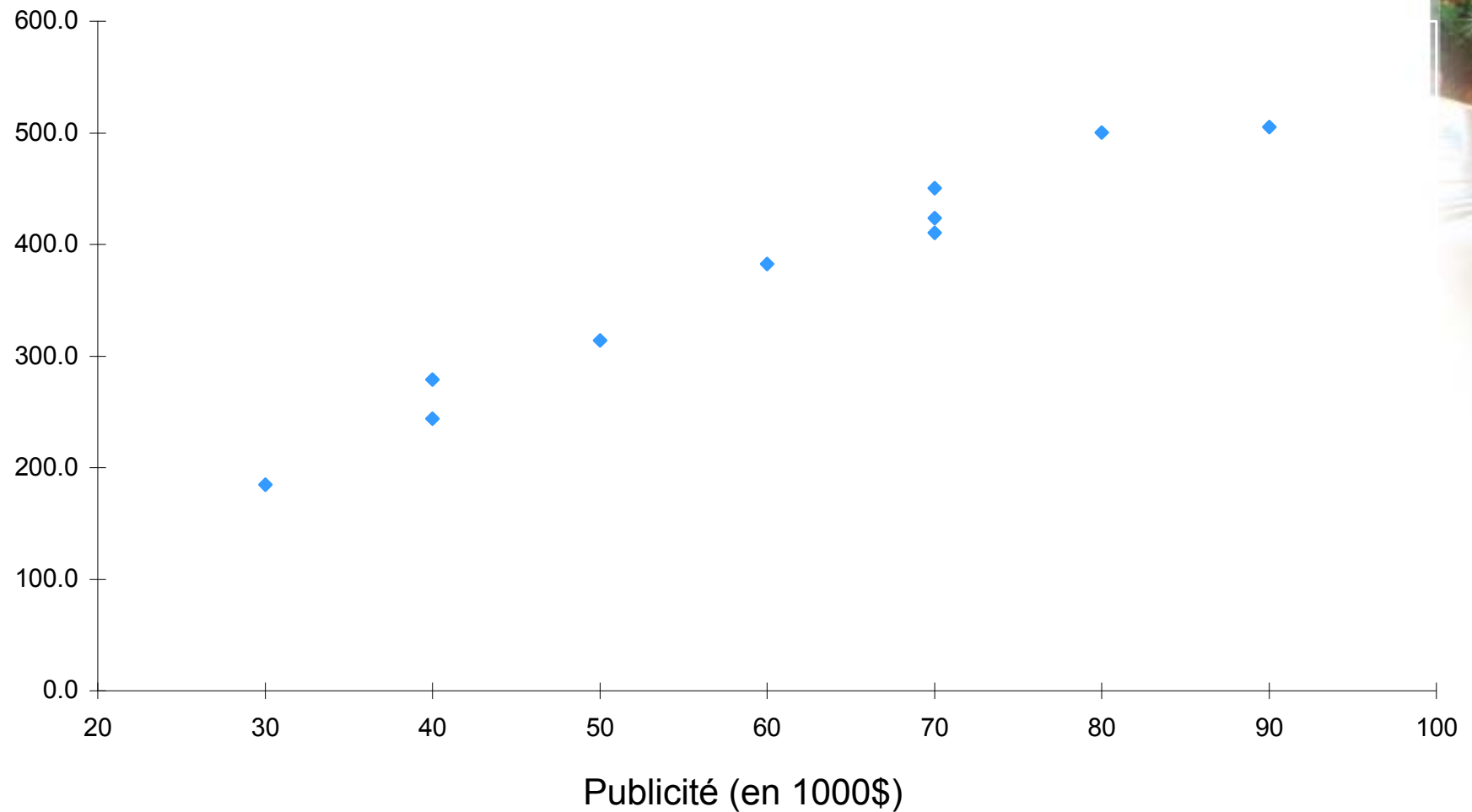
**X**

## 5.1 Un exemple

- Considérons la relation entre le montant investi en publicité ( $X_1$ ) et le niveau des ventes ( $Y$ ) pour une compagnie. Voir le fichier [4600-5.1-Ventes vs Pub-Data.xls](#) pour les données.
- Il existe probablement une relation entre ces deux variables (si le montant en publicité augmente, les ventes devraient augmenter).
- Comment mesurer et quantifier cette relation?

## 5.1 Graphique de dispersion

Ventes (en 1000\$)





## 5.2 Régression linéaire simple

- Le graphique semble indiquer une relation linéaire entre le montant en publicité et le niveau des ventes.
- Ainsi, un modèle de régression linéaire semble approprié pour représenter la vraie relation (pour la population entière) entre la publicité et les ventes.

$$Y_i = \beta_0 + \beta_1 X_{1_i} + \varepsilon_i$$

- Comment interpréter les paramètres?
  - $\beta_1$  = pente de la droite  
= effet de la variable explicative sur la moyenne de Y  
(Toutes autres choses étant égales par ailleurs)
  - $\beta_0$  = ordonnée à l'origine  
= valeur de la moyenne de Y quand X est nul (si applicable)

## 5.2 Régression linéaire simple

- Le modèle de régression estimé (basée sur notre échantillon) va être représenté par

$$\hat{Y}_i = b_0 + b_1 X_{1_i}$$

$\hat{Y}_i$  est la valeur estimée de Y à un niveau donné de X

- Des valeurs numériques doivent être assignées à  $b_0$  et  $b_1$ .
- Comment les choisir?



## 5.2 Estimation des paramètres

- La méthode des “moindre carrés” choisit les valeurs qui minimise :

$$SCE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - (b_0 + b_1 X_{1_i}))^2$$

- Si  $SCE = 0$  notre fonction estimée passe par tous les points (ajustement parfait).
- Pour l'exemple, on obtient :

$$\hat{Y}_i = 36.342 + 5.550X_{1_i}$$

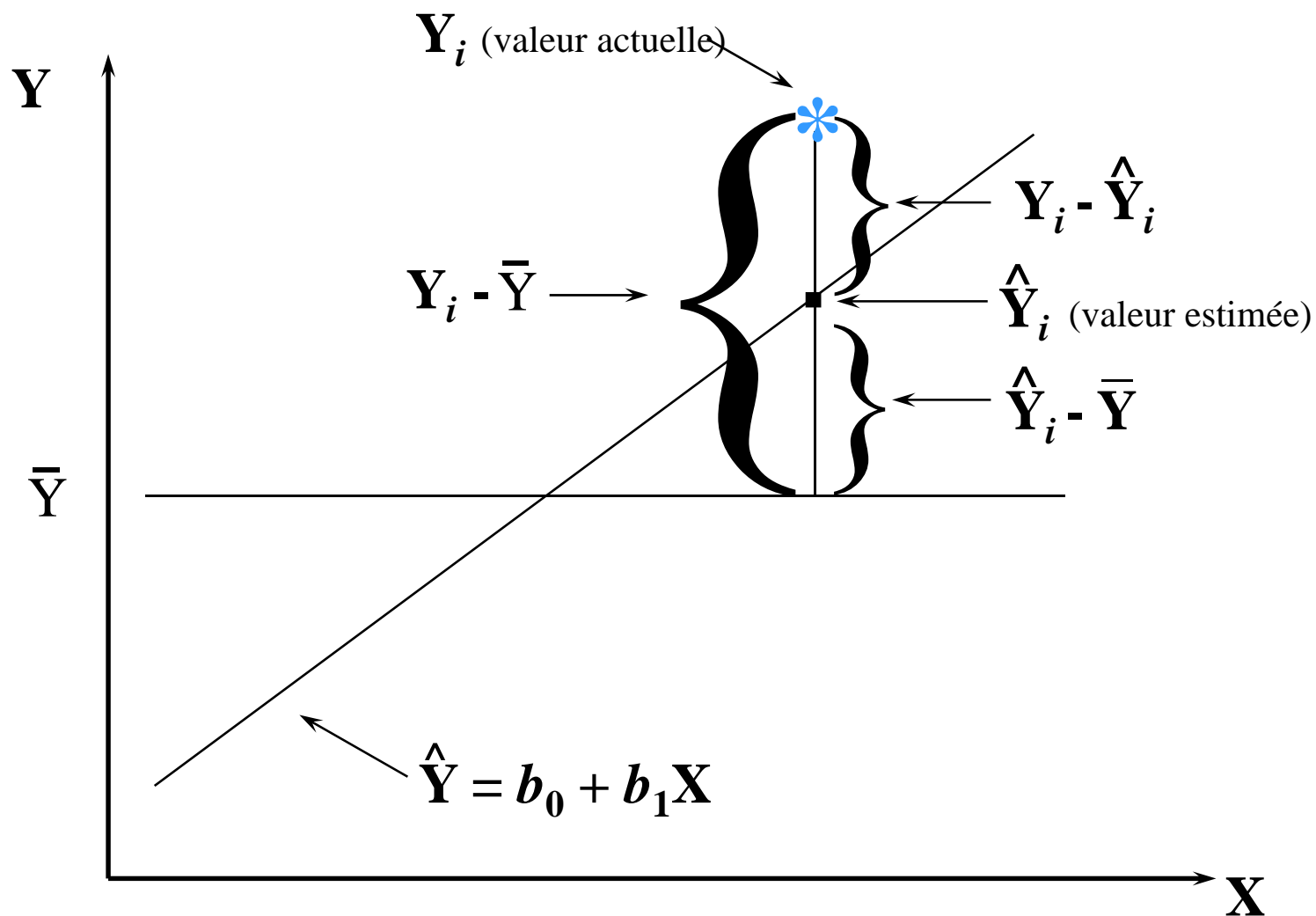
## 5.2 Estimation des paramètres avec Excel

1. Par l'utilitaire d'analyse (voir Ragsdale, section 9.6). On obtient alors un rapport détaillée.
2. Par la courbe de tendance (voir Ragsdale, section 9.7). On obtient alors l'équation, le  $R^2$  ainsi que la droite.
3. Par la fonction « tendance » (voir Ragsdale, section 9.7). Les paramètres ne sont pas affichés mais on peut calculer des estimations. Cette fonction sera utilisée à la séance 6.

## 5.2 Coefficient de détermination

- Le coefficient de détermination ( $R^2$ ) est un indicateur de la qualité de l'ajustement de la droite de régression aux données.
- $0 \leq R^2 \leq 1$
- Il mesure la proportion de la variation totale de Y autour de sa moyenne qui est expliquée par le modèle de régression.
- Disponible dans le rapport de régression offert par l'utilitaire d'analyse d'Excel (2<sup>e</sup> ligne dans le premier tableau du rapport – statistiques de la régression).
- Pour notre exemple, on obtient une valeur de 96.9.1%.

## 5.2 Décomposition de l'erreur



## 5.2 Partition de la somme des carrés

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

ou,

$$\text{SCT} = \text{SCE} + \text{SCR}$$

$$R^2 = \frac{\text{SCR}}{\text{SCT}} = 1 - \frac{\text{SCE}}{\text{SCT}}$$

## 5.2 Test t

- ❑ Permet de tester que chaque paramètre de la droite de régression est significativement différent de 0 (relation linéaire entre les deux variables).
  - $H_0 : \beta_1 = 0$  (vraie pente est nulle)
  - $H_1 : \beta_1 \neq 0$  (vraie pente est non nulle)
- ❑ On rejette  $H_0$  si le seuil expérimental (p-value) est inférieur au seuil de signification (généralement, 5%). Le seuil expérimental est donné par l'utilitaire d'analyse d'Excel (5e colonne dans le troisième tableau du rapport).
- ❑ Pour notre exemple, le seuil expérimental est de  $2.51 \times 10^{-7}$ , ce qui est nettement inférieur à 5%. Ainsi, on conclut qu'il existe une relation entre les deux variables.



## 5.2 Analyse de variance (ANOVA)

- Permet de tester la signification de la RLS en déterminant la part de la variation de Y expliquée par la RLS et celle non expliquée par RLS (résidu).
- Résidu indique pour chaque observation, la différence entre la valeur réelle de Y et celle estimée par le modèle.
- La dernière colonne donne la statistique F. Pour une régression linéaire simple, c'est la même chose que le test t mais pour une régression linéaire multiple, on testerait que la régression est globalement significative :
  - H0 : Toutes les pentes sont nulles
  - H1 : Au moins une pente est non nulle

## 5.2 Erreur-type

- L'erreur-type (noté  $S_e$ ) mesure la dispersion dans les données autour de la droite de régression estimée.

$$S_e = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - k - 1}}$$

où  $k$  = nombre de variables indépendantes

- Pour notre exemple,  $S_e = 20.421$
- Cette valeur est utile pour faire des prédictions (voir page 19)...

## 5.2 Inférence et extrapolation

- La régression peut être utilisée pour estimer de façon ponctuelle ou par un intervalle :
  - la moyenne conditionnelle (moyenne de  $Y$  pour une valeur de  $X$ );
  - la valeur de  $Y$  pour une valeur de  $X$  (voir page suivante).

Note : une extrapolation à l'extérieur du domaine de l'échantillon est dangereuse.

## 5.2 Intervalle de prédiction

- Une approximation de l'intervalle à 95% pour une prédiction d'une nouvelle valeur de Y lorsque  $X_1 = X_{1h}$  est donné par

Où :

$$\hat{Y}_h \pm 2S_e$$

$$\hat{Y}_h = b_0 + b_1 X_{1h}$$

- ◆ Exemple : Si 65 000\$ est dépensé en publicité :  
Borne inférieure =  $397.092 - 2 \times 20.421 = 356.250$   
Borne supérieure =  $397.092 + 2 \times 20.421 = 437.934$
- ◆ Si nous dépensons 65 000\$ en publicité, nous sommes approximativement confiant à 95% que les ventes vont se situer entre 356 250\$ et 437 934\$.

## 5.2 Hypothèses du modèle

- **La relation est linéaire** : peut se vérifier en examinant le nuage de points des données (relation linéaire) ou le graphe des résidus (pas de relation précise).
- **Les résidus sont normaux** : peut se vérifier en examinant la distribution des résidus pour chaque valeur de  $X$  (graphiquement ou test d'ajustement vu à la séance 4).
- **La variance des résidus est constante** : peut se vérifier en regardant s'il y a d'énormes différences dans le graphe des résidus selon la valeur de  $X$ .
- **Les résidus ne sont pas corrélés** : peut se vérifier à l'aide de la statistique Durbin-Watson (disponible dans certains logiciels).
- Voir la section 9.10.2 de Ragsdale pour plus de détails.

## 5.3 Régression linéaire multiple (RLM)

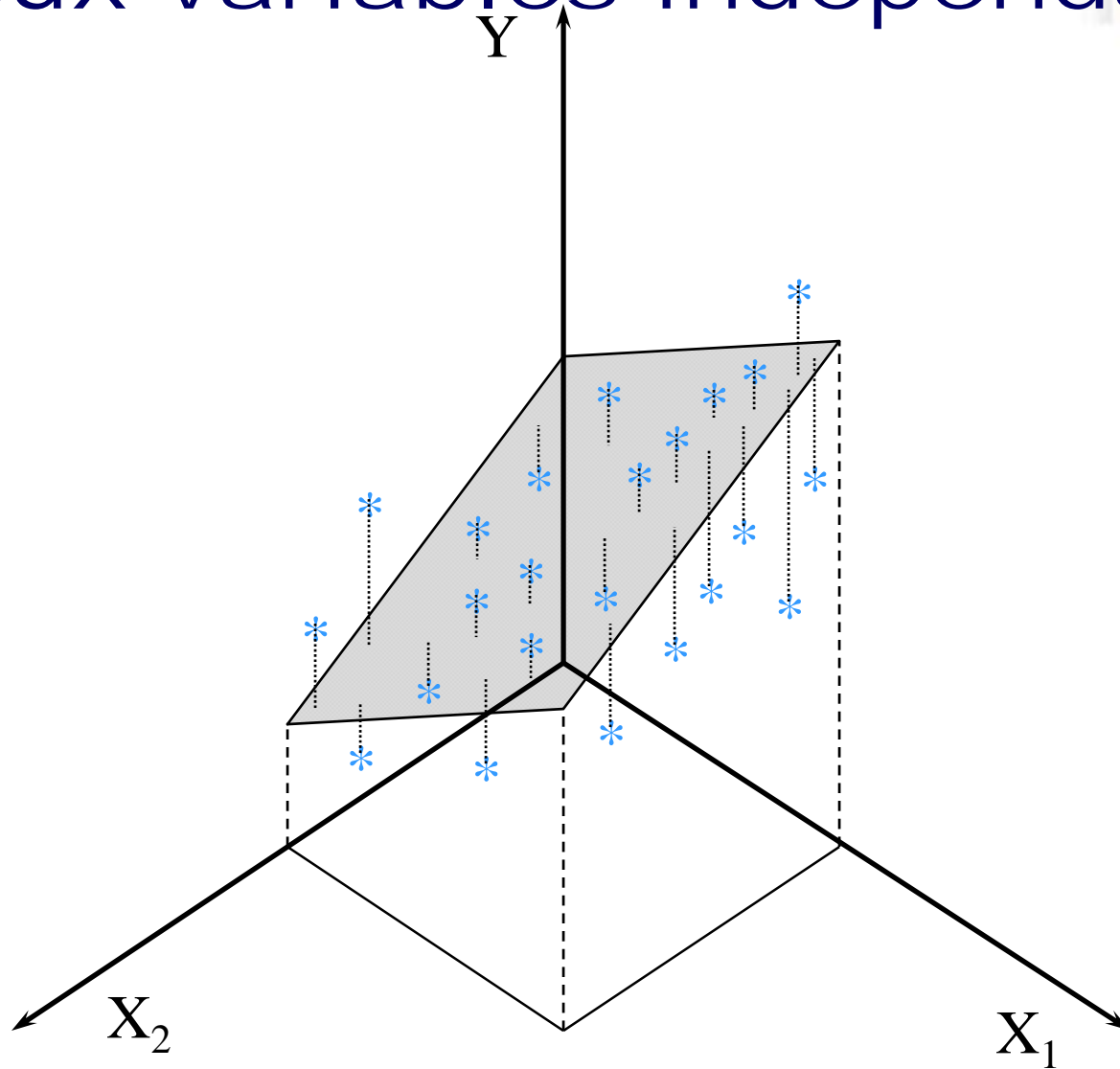
- La plupart des modèles de régression implique plus d'une variable indépendante.
- Si chaque variable indépendante varie de façon linéaire avec Y, le modèle estimé de régression devient alors :

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i} + \dots + b_k X_{k_i}$$

- Le coefficient représente l'effet de la variable explicative sur la moyenne de Y (*toutes autres choses étant égales par ailleurs*).
- Les valeurs optimales des  $b_i$  peuvent encore être obtenues en minimisant la SCE.
- La fonction résultante ajuste un hyperplan à notre échantillon de données.



## 5.3 Deux variables indépendantes



## 5.3 RLM avec Excel

- ❑ L'utilitaire d'analyse permet de faire un modèle de RLM.
- ❑ Il faut alors s'assurer que toutes les variables indépendantes se retrouvent dans des colonnes adjacentes.

## 5.3 Exemple

- Un évaluateur foncier aimerait développer un modèle pour l'aider à prédire le prix de vente de propriétés résidentielles.
- Trois variables indépendantes vont être utilisées pour estimer le prix de vente d'une habitation :
  - Superficie totale ( $X_1$ )
  - Taille du garage ( $X_2$ )
  - Nombre de chambres ( $X_3$ )
- Voir les données dans [4600-5.3-Prix\\_Maison-Data.xls](#).

## 5.4 Choix des variables explicatives

- Nous voulons trouver le modèle le plus simple qui explique adéquatement la variation systématique de  $Y$ .
- Utiliser arbitrairement toutes les variables indépendantes peut résulter en un sur-ajustement.
- Un échantillon reflète des caractéristiques :
  - Représentatives de la population
  - Spécifiques à l'échantillon
- Nous voulons éviter l'ajustement de caractéristiques spécifiques de l'échantillon (sur-ajustement des données).

## 5.4 Modèles avec une variable indépendante

- Pour simplifier, supposons qu'on ajuste trois modèles de régression linéaire simple :

$$\hat{Y}_i = b_0 + b_1 X_{1i}$$

$$\hat{Y}_i = b_0 + b_2 X_{2i}$$

$$\hat{Y}_i = b_0 + b_3 X_{3i}$$

- Les résultats sommaires sont :

Variables dans le modèle	R <sup>2</sup>	R <sup>2</sup> -ajusté	S <sub>e</sub>	Estimation des paramètres
X <sub>1</sub>	0.870	0.855	10.299	b <sub>0</sub> =9.503, b <sub>1</sub> =56.394
X <sub>2</sub>	0.759	0.731	14.030	b <sub>0</sub> =78.290, b <sub>2</sub> =28.382
X <sub>3</sub>	0.793	0.770	12.982	b <sub>0</sub> =16.250, b <sub>3</sub> =27.607

- Le modèle utilisant X<sub>1</sub> explique 87% de la variation de Y (13% inexpliqué).

## 5.4 Modèles avec deux variables indépendantes

- Maintenant, supposons qu'on désire étendre le modèle à deux variables indépendantes :

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i}$$

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_3 X_{3_i}$$

- ◆ Les résultats sommaires sont :

Variables dans le modèle	R <sup>2</sup>	R <sup>2</sup> -ajusté	S <sub>e</sub>	Estimation des paramètres
X <sub>1</sub>	0.870	0.855	10.299	b <sub>0</sub> =9.503, b <sub>1</sub> =56.394
X <sub>1</sub> & X <sub>2</sub>	0.939	0.924	7.471	b <sub>0</sub> =27.684, b <sub>1</sub> =38.576 b <sub>2</sub> =12.875
X <sub>1</sub> & X <sub>3</sub>	0.877	0.847	10.609	b <sub>0</sub> =8.311, b <sub>1</sub> =44.313 b <sub>3</sub> =6.743

- ◆ Le modèle utilisant X<sub>1</sub> et X<sub>2</sub> explique 93.9% de la variation de Y.



## 5.4 $R^2$ -ajusté

- L'ajout de variables indépendantes à un modèle :
  - Augmente automatiquement la valeur de  $R^2$
  - Le  $R^2$ -ajusté peut cependant diminuer ou augmenter.

$$R_a^2 = 1 - \left( \frac{\text{SCE}}{\text{SCT}} \right) \left( \frac{n-1}{n-k-1} \right)$$

- Nous pouvons comparer les valeurs du  $R^2$ -ajusté pour vérifier heuristiquement si l'ajout d'une variable indépendante a vraiment permis d'améliorer le modèle de régression.

## 5.4 Modèle avec trois variables indépendantes

- Supposons qu'on utilise les trois variables :

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i} + b_3 X_{3_i}$$

- ◆ Les résultats sommaires sont :

Variables dans le modèle	R <sup>2</sup>	R <sup>2</sup> -ajusté	S <sub>e</sub>	Estimation des paramètres
X <sub>1</sub>	0.870	0.855	10.299	b <sub>0</sub> =9.503, b <sub>1</sub> =56.394
X <sub>1</sub> & X <sub>2</sub>	0.939	0.924	7.471	b <sub>0</sub> =27.684, b <sub>1</sub> =38.576, b <sub>2</sub> =12.875
X <sub>1</sub> , X <sub>2</sub> & X <sub>3</sub>	0.943	0.918	7.762	b <sub>0</sub> =26.440, b <sub>1</sub> =30.803, b <sub>2</sub> =12.567, b <sub>3</sub> =4.576

- ◆ Le modèle avec X<sub>1</sub> et X<sub>2</sub> semble être le meilleur :
  - R<sup>2</sup>-ajusté le plus élevé
  - S<sub>e</sub> le plus faible (intervalles de prédiction plus précis)

## 5.4 Remarque sur la multicollinéarité

- Ce n'est pas surprenant que l'ajout de  $X_3$  (# de chambres) au modèle avec  $X_1$  (superficie totale) n'améliore pas significativement le modèle.
- Les deux variables représentent la même chose (ou à peu près), i.e., la taille de la maison.
- Ces deux variables sont donc fortement auto-corrélées (ou colinéaires).

## 5.4 Prédiction

- Estimons le prix de vente moyen d'une maison de 2 100 pieds-carré avec un garage double :

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i}$$

$$\hat{Y}_i = 27.684 + 38.576 * 2.1 + 12.875 * 2 = 134.444$$

- ◆ Le prix moyen estimé est de 134 444\$.
- ◆ Un intervalle à 95% sur le prix d'une maison donne approximativement :

$$\hat{Y}_h \pm 2S_e$$

borne inférieure =  $134.444 - 2 * 7.471 = \$119,502$

borne supérieure =  $134.444 + 2 * 7.471 = \$149,386$

## 5.5 Variables binaires indépendantes

- D'autres types de facteurs non quantitatifs peuvent être considérés comme variables indépendantes à l'aide de variables binaires.
- ◆ Exemple : La présence (ou absence) d'une piscine,

$$X_{p_i} = \begin{cases} 1, & \text{si la maison } i \text{ possède une piscine} \\ 0, & \text{sinon} \end{cases}$$

## 5.5 Variables binaires indépendantes

- ◆ Exemple : La condition du toit (bonne, moyenne ou mauvaise)
- ◆ Voir le fichier 4600-5.5-Prix\_Maison-Sol.xls

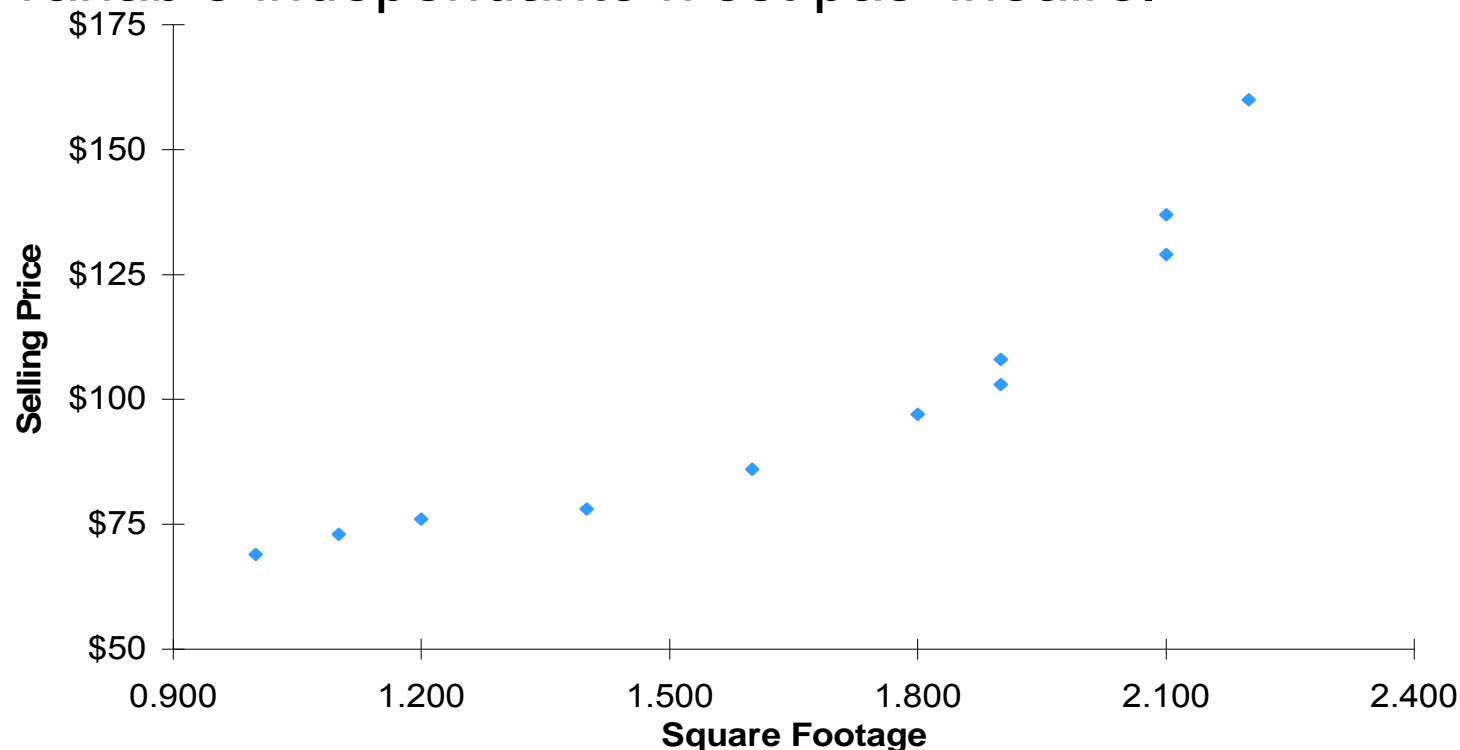
$$X_{4_i} = \begin{cases} 1, & \text{si le toit de la maison } i \text{ est en bonne condition} \\ 0, & \text{sinon} \end{cases}$$

$$X_{5_i} = \begin{cases} 1, & \text{si le toit de la maison } i \text{ est en moyenne condition} \\ 0, & \text{sinon} \end{cases}$$



## 5.6 Relations non linéaires

- Parfois la relation entre une variable dépendante et une variable indépendante n'est pas linéaire.



- ◆ Ce graphique suggère une relation quadratique entre la superficie totale (X) et le prix de vente (Y).

## 5.6 Modèle de régression quadratique

- Un modèle approprié de régression dans ce cas peut être,

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{1_i}^2$$

Ou de façon équivalente,

$$\hat{Y}_i = b_0 + b_1 X_{1_i} + b_2 X_{2_i}$$

avec,

$$X_{2_i} = X_{1_i}^2$$

## 5.6 Régression quadratique avec Excel

1. Ajouter une colonne dans le chiffrier avec le calcul du carré de la variable. Faire ensuite une RLM avec l'utilitaire d'analyse.
2. Ajouter une courbe de tendance (polynomiale de degré 2) sur un graphique de dispersion.

## 5.6 Graphique de la courbe estimée de régression

