



Analyse de Données

Par: Houda Benbrahim



Plan du Cours

- Chapitre I: Rappel: Statistique Descriptive
- Chapitre II: Régression Linéaire Simple et Multiple
- Chapitre III: Analyse Factorielle: ACP
- Chapitre IV: Techniques de Classifications
- En parallèle: Travaux Pratiques



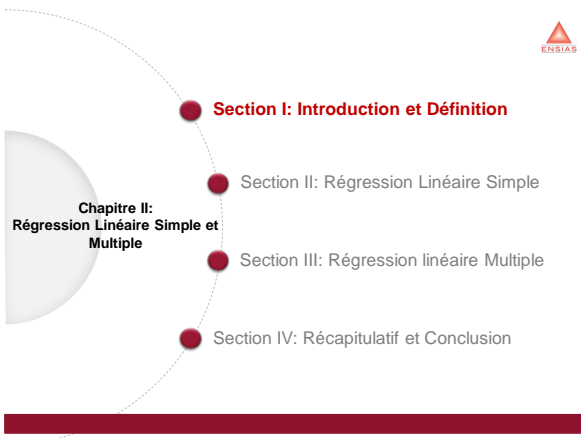
Plan du Cours

- Chapitre I: Rappel: Statistique Descriptive
- **Chapitre II: Régression Linéaire Simple et Multiple**
- Chapitre III: Analyse Factorielle: ACP
- Chapitre IV: Techniques de Classifications
- En parallèle: Travaux Pratiques



Chapitre II: Régression Linéaire Simple et Multiple

- Section I: Introduction et Définition
- Section II: Régression Linéaire Simple
- Section III: Régression linéaire Multiple
- Section IV: Récapitulatif et Conclusion



Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

Contexte général Définition Applications

Exemple:

Supposant qu'on veut acheter un appartement à Rabat. Une recherche dans les sites immobiliers nous a conduit aux données suivantes:

Quartier	Superficie (m²)	Prix de Vente (Dhs)
Agdal	135	2 000 000
Hay Ryad	257	5 140 000
Hassan	180	3 400 000
Agdal	120	2 000 000
Qbibat	73	850 000

H. Benbrahim

Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

Contexte général Définition Applications

Exemple:

Table de données:

Quartier	Superficie m²	Etage	HS	Prix de Vente Dhs
Agdal	135	1	Non	2 000 000
Hay Ryad	257	3	Oui	5 140 000
Hassan	180	5	Oui	3 400 000
Agdal	120	3	Oui	2 000 000
Qbibat	73	2	Non	850 000
Agdal				
Guich				
Hay Ryad				
Agdal				

H. Benbrahim

Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

Contexte général Définition Applications

Exemple: Questions à se poser?

- Comment relier les prix de vente à la superficie des appartements?
- Quelle est le prix de vente espérés si la superficie qu'on désire est 200 m² ?
- Est-ce que la surface d'un appartement détermine assez largement son prix?
 - pas complètement !
 - autres facteurs à prendre en compte
 - ✓ quartier
 - ✓ étage
 - ✓ Haut standing
 - ✓ présence ascenseur, orientation, parking, gardien, année de construction, etc.

H. Benbrahim

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications	

Exemple: Solution:

- Analyser les **données** afin d'en dégager des informations nouvelles qui vont fonder des décisions.
- Prendre des décisions

9 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications	

Exemple: Etude statistique:

- On distingue 2 objectifs:
- On cherche à savoir s'il existe un lien entre *la superficie et le prix de vente*.
- On cherche à savoir si *la superficie a une influence sur le prix de vente et éventuellement prédire le prix de vente à partir de la superficie*.

10 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications	

Exemple: Etude statistique:

- On distingue 2 objectifs:
- On cherche à savoir s'il existe un lien entre *la superficie et le prix de vente*.
 - Liaison entre les variables . On définit un indice de liaison : coeff. De corrélation, statistique du Khi-2,...
- On cherche à savoir si *la superficie a une influence sur le prix de vente et éventuellement prédire le prix de vente à partir de la superficie*.
 - Influence de variables sur une autre. On modélise cette influence: régression logistique, **régression linéaire**,...

11 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications	

Régression: Définition

- La régression est une technique de modélisation qui permet de mettre en équation une relation entre une variable endogène (à expliquer) et n variables exogènes (explicatives).
- La régression permet d'analyser la manière dont une variable (dite expliquée) est affectée par les valeurs d'une ou plusieurs autres variables (dites explicatives)
- Un problème de régression consiste à chercher une fonction f telle que pour tout i , Y_i soit approximativement égale à $f(X_i)$.

12 H. Benbrahim ENSIAS

Définition		
Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications
Régression: Définition <ul style="list-style-type: none"> Un problème de régression consiste à chercher une fonction f telle que pour tout i, Y_i soit approximativement égale à $f(X_i)$. 		
Variable à expliquer	Variables explicatives	Nom de l'analyse
1 quantitative	1 quantitative	régression simple
1 quantitative	plusieurs quantitatives	régression multiple
1 quantitative	plusieurs qualitatives	analyse de variance
1 quantitative	plusieurs quantitatives et qualitatives	analyse de covariance
1 qualitative	plusieurs quantitatives et qualitatives	régression logistique
1 qualitative	plusieurs quantitatives	analyse discriminante probabiliste

13

H. Benbrahim



Définition		
Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications
Régression Linéaire: Définition <ul style="list-style-type: none"> Cette technique est couramment utilisée lorsque l'on souhaite prédire la réalisation d'une variable de type continue à l'aide d'un ensemble de variables, dits prédicteurs, du même type. La régression linéaire permet de modéliser une relation entre une variable endogène (ou dépendante) Y et p variables exogènes (ou indépendantes) X_1, X_2, \dots, X_p: $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$ <ul style="list-style-type: none"> Les β_i sont les coefficients que l'on va chercher à estimer. ε est la partie aléatoire que l'on ne peut contrôler. On l'appelle aussi erreur. 		

14

H. Benbrahim



Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Contexte général

Définition

Applications

Régression Linéaire: Exemple

* On veut représenter la *consommation* d'un agent énergétique en fonction de facteurs explicatifs :

- La température moyenne sur un mois d'un ménage
- L'épaisseur de l'isolation du logement

	Cosommation Gallon/mois	Isolation (en cm)	Température Moyenne (°F)
1	275,30	3,00	40,00
2	363,80	3,00	27,00
3	164,30	10,00	40,00
4	40,80	6,00	73,00
5	94,30	6,00	64,00
6	230,90	6,00	34,00
7	366,70	6,00	9,00
8	300,60	10,00	8,00
9	237,80	10,00	23,00
10	121,40	3,00	63,00
11	31,40	10,00	65,00
12	203,50	6,00	41,00
13	441,10	3,00	21,00
14	323,00	3,00	38,00
15	52,50	10,00	58,00

15

ENSIAS

15



Définition		
Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications
Régression Linéaire: Exemple <ul style="list-style-type: none"> <i>consommation</i> énergétique en fonction de la température moyenne mensuelle et de l'épaisseur de l'isolation du logement. 		
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$ <div style="display: flex; justify-content: space-around; align-items: flex-start;"> <div style="text-align: center;"> <p>Observation i de la Consommation mensuelle</p> </div> <div style="text-align: center;"> <p>Terme constant</p> </div> <div style="text-align: center;"> <p>Influence de l'isolation</p> </div> <div style="text-align: center;"> <p>Influence de la Température</p> </div> <div style="text-align: center;"> <p>Erreur aléatoire</p> </div> </div>		

16

H. Benbrahim



Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications	

Régression Linéaire: Applications

- Le modèle de régression linéaire a de nombreuses applications pratiques.
- Il permet de faire des analyses de prédiction:
 - estimer un modèle de régression linéaire
 - prédire quel serait le niveau de y pour des valeurs particulières de x .
- Il permet d'estimer l'effet d'une variable sur une autre contrôlée par d'autres facteurs.
 - dans le domaine des sciences de l'éducation, on peut évaluer l'effet de la taille des classes sur les performances scolaires des enfants contrôlée par la catégorie socioprofessionnelle des parents ou par l'emplacement géographique de l'établissement.

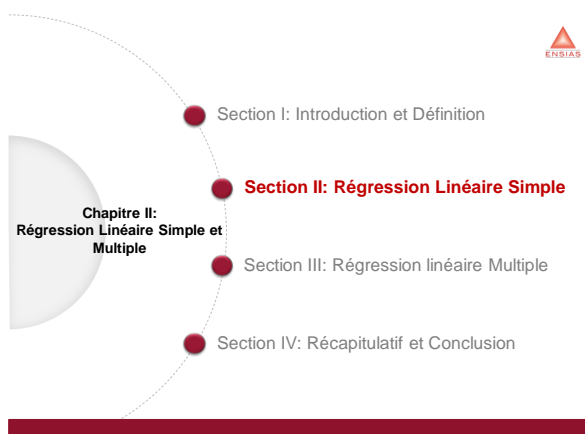
17 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Contexte général	Définition	Applications	

Régression Linéaire: Applications

- Econométrie:
 - Modèle linéaire pour estimer l'effet du nombre de policiers sur la criminalité.
 - Régression linéaire pour estimer l'effet des institutions sur le développement actuel des pays.
 - Modèle linéaire pour analyser sur des données américaines l'effet des lois autorisant le travail le dimanche sur la participation religieuse.
- Sociologie:
 - La structure sociale européenne est analysée à l'aide de la régression linéaire entre l'écart type du niveau de revenu et celui du niveau d'éducation.
 - La régression linéaire pour évaluer l'estime de soi en fonction du niveau de consommation de cannabis, de l'âge et du sexe.
- ...

18 H. Benbrahim ENSIAS



Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence

Régression Linéaire Simple

- La relation entre deux variables x et y est décrite par:

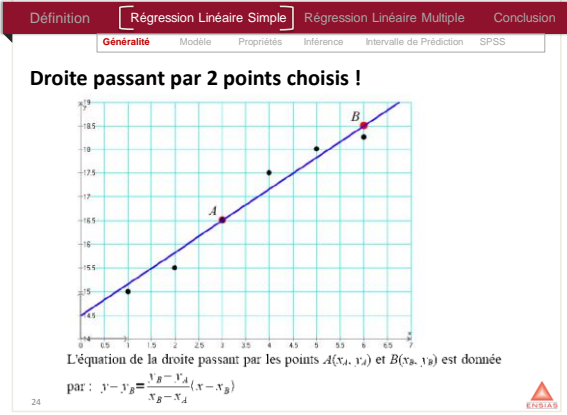
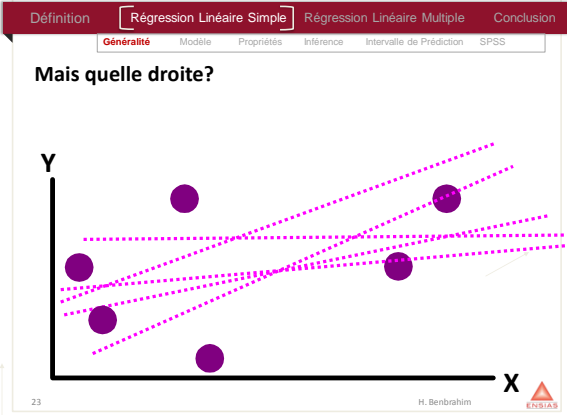
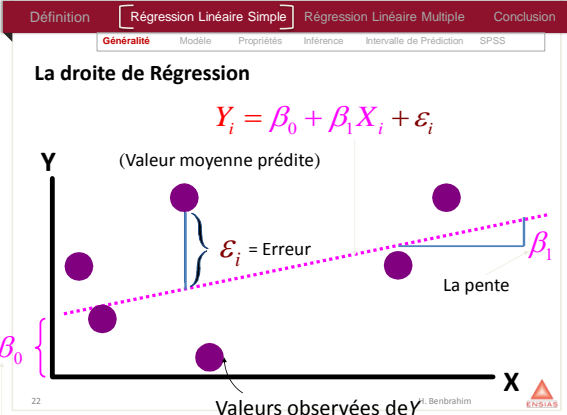
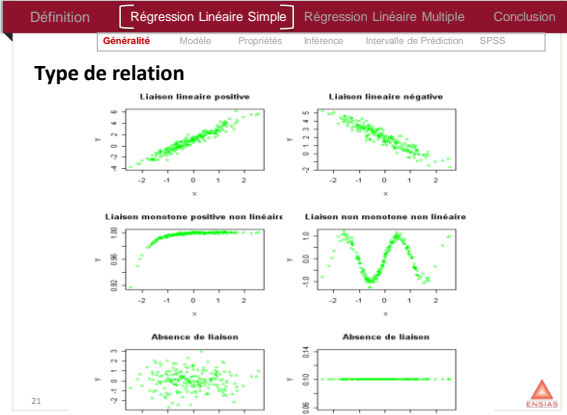
$$y = \beta_0 + \beta_1 x + \varepsilon$$

Où β_0 et β_1 sont deux constantes que l'on cherche à évaluer et ε est un terme aléatoire que l'on appelle erreur.

Pour estimer β_0 et β_1 on dispose d'un échantillon $(x_1, y_1), \dots, (x_n, y_n)$ supposé vérifier:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ pour } i = 1, 2, \dots, n.$$


20 H. Benbrahim ENSIAS



Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Droite de Mayer



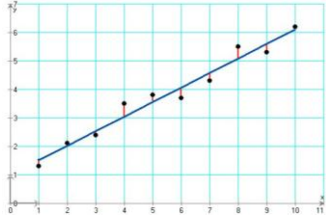
On découpe le nuage de points en deux sous-ensembles de même effectif. Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et la moyenne des y_i . On obtient ainsi deux points (\bar{x}_1, \bar{y}_1) et (\bar{x}_2, \bar{y}_2) , appelés **points moyens**. Il reste à tracer la droite passant par ces deux points.

25 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Droite par la méthode des moindres carrés



L'ajustement linéaire par la méthode des moindres carrés consiste à déterminer la droite (que l'on appelle aussi **droite de régression**) telle que la somme des carrés des n valeurs $y_i - \hat{y}_i$ soit minimale (ce qui explique le nom de la méthode).

26 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Modèle de Régression Linéaire Simple

- La relation entre deux variables x et y est décrite par:

$$y_i = a \times x_i + b + \varepsilon_i, \text{ pour } i = 1, 2, \dots, n.$$
 - a et b sont les paramètres du modèle.
 - a est la pente, b est la constante.
 - ε est l'erreur du modèle.
 - ε résume toute l'information qui n'est pas prise en compte dans la relation linéaire.
 - Les propriétés des estimateurs reposent sur les hypothèses que nous formulons sur ε .

27 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Exemple: Rendement de maïs et quantité d'engrais

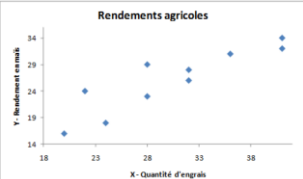
- On dispose de $n = 10$ observations. On cherche à expliquer Y le rendement en maïs (en quintal) de parcelles de terrain, à partir de X la quantité d'engrais (en kg) que l'on y a épandu.
- L'objectif est de modéliser le lien à travers une relation linéaire.
- si l'on ne met pas d'engrais du tout, il sera quand même possible d'obtenir du maïs, c'est le sens de la constante b de la régression. Sa valeur devrait être positive.
- Ensuite, plus on mettra de l'engrais, meilleur sera le rendement. On suppose que cette relation est linéaire, d'où l'expression $a \times x$, on imagine à l'avance que a devrait être positif.

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Exemple: Rendement de maïs et quantité d'engrais

- Le graphique nuage de points associant X et Y semble confirmer cette première analyse.
- Dans le cas contraire où les coefficients estimés contredisent les valeurs attendues (b ou/et a sont négatifs):
 - une perception faussée du problème,
 - les données utilisées ne sont pas représentatives du phénomène que l'on cherche à mettre en exergue,
 - ou bien...



29

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Hypothèses (1/3)

- Ces hypothèses pèsent sur les propriétés des estimateurs (biais, convergence) et l'inférence statistique (distribution des coefficients estimés).

H1 - Hypothèses sur Y et X .

- ✓ X et Y sont des grandeurs numériques mesurées sans erreur.
- ✓ X est une donnée exogène dans le modèle. Elle est supposée non aléatoire.
- ✓ Y est aléatoire par l'intermédiaire de ϵ i.e. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs dans le modèle.

30 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Hypothèses (2/3)

H2 - Hypothèses sur le terme aléatoire ϵ .

- ✓ Les ϵ_i sont i.i.d (indépendants et identiquement distribués).

H2.a $E(\epsilon_i) = 0$

- en moyenne les erreurs s'annulent c.-à-d. le modèle est bien spécifié.

H2.b $V(\epsilon_i) = \sigma^2$

- la variance de l'erreur est constante et ne dépend pas de l'observation.

C'est l'hypothèse d'homoscédasticité.

H2.c $COV(x_i, \epsilon_i) = 0$

- l'erreur est indépendante de la variable exogène

31 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Hypothèse (3/3)

H2.d $COV(\epsilon_i, \epsilon_j) = 0$

- Indépendance des erreurs.
- Les erreurs relatives à 2 observations sont indépendantes.
- On parle de "non auto-corrélation des erreurs".

H2.e $\epsilon_i \equiv N(0, \sigma)$.

- L'hypothèse de normalité des erreurs est un élément clé pour l'inférence statistique.

32 H. Benbrahim ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

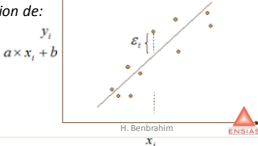
Intervalle de Prédiction

SPSS

Estimateur des moindres carrés

- Objectif = déterminer les valeurs de a et b en utilisant les informations apportées par l'échantillon.
 - estimation meilleure → la droite de régression doit approcher au mieux le nuage de points.
- Le critère des moindres carrés → minimiser la somme des carrés des écarts (des erreurs) entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction.
- L'estimateur des moindres carrés des paramètres a et b répond à la minimisation de:

$$S = \sum_{i=1}^n e_i^2$$
$$= \sum_{i=1}^n [y_i - (ax_i + b)]^2$$
$$= \sum_{i=1}^n [y_i - ax_i - b]^2$$



H. Benbrahim

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Estimateur des moindres carrés

- S minimum → dérivée première par rapport à a et b sont 0.

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

H. Benbrahim

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Estimateur des moindres carrés

- Avec les dérivées partielles de S par rapport à a et à b , on obtient le système:

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \sum_{i=1}^n (y_i - a \cdot x_i - b) = 0 \\ \sum_{i=1}^n x_i (y_i - a \cdot x_i - b) = 0 \end{cases}$$
$$\Rightarrow \begin{cases} \sum_{i=1}^n y_i - a \sum_{i=1}^n x_i - b \sum_{i=1}^n 1 = 0 \\ \sum_{i=1}^n x_i y_i - a \sum_{i=1}^n x_i^2 - b \sum_{i=1}^n x_i = 0 \end{cases}$$
$$\Rightarrow \begin{cases} \sum_{i=1}^n y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n 1 \\ \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^n y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n 1 \\ \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i \end{cases}$$

H. Benbrahim

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Estimateur des moindres carrés

En appelant \hat{a} et \hat{b} les solutions de ces équations normales, nous obtenons les estimateurs des moindres carrés :

$$\hat{a} = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \cdot \bar{y} \cdot n}{\sum_{i=1}^n x_i^2 - \bar{x}^2 \cdot n}$$
$$\hat{b} = \frac{\sum_{i=1}^n y_i - \hat{a} \cdot \sum_{i=1}^n x_i}{n}$$

La droite passe donc par le point moyen G de coordonnées : (\bar{X}, \bar{Y})

H. Benbrahim

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Exemple: Rendement Agricole

i	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB) ²
1	16	20	-10.1	-10.4	105.04	108.16
2	18	24	-8.1	-6.4	51.84	40.96
3	23	28	-3.1	-2.4	7.44	5.76
4	24	22	-2.1	-8.4	17.64	70.56
5	28	32	1.9	1.6	3.04	2.56
6	29	26	2.9	-2.4	-6.96	5.76
7	26	32	-0.1	1.6	-0.16	2.56
8	31	36	4.9	5.6	27.44	31.36
9	32	41	5.9	10.6	62.54	112.36
10	34	41	7.9	10.6	83.74	112.36
Moyenne	26.1	30.4				
			Somme	351.6	492.4	
			a*	0.7141		
			b*	4.3928		

- Nous calculons les moyennes des variables, $\bar{y} = 26.1$ et $\bar{x} = 30.4$.
- Nous formons alors les valeurs de $(y_i - \bar{y})$, $(x_i - \bar{x})$, $(y_i - \bar{y}) \times (x_i - \bar{x})$ et $(x_i - \bar{x})^2$.
- Nous réalisons les sommes $\sum_i (y_i - \bar{y}) \times (x_i - \bar{x}) = 351.6$ et $\sum_i (x_i - \bar{x})^2 = 492.4$.

37 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Exemple: Rendement Agricole

les estimations :

$$\hat{a} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{351.6}{492.4} = 0.7141$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 26.1 - 0.7141 \times 30.4 = 4.3928$$

38 H. Benbrahim ENSIAS

Droite de régression - "Rendements agricoles"

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Exemple: Rendement Agricole

- On constate que la droite passe plus ou moins au milieu du nuage de points.
- Est-ce que notre modélisation est suffisamment intéressante?
 - Evaluation visuelle ➔ ne suffit pas.
 - Critère quantitatif?

39 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Erreur et Résidu

- ϵ est l'erreur inconnue introduite dans la spécification du modèle.
- la valeur prédite de l'endogène Y pour l'individu i: $\hat{y}_i = \hat{y}(x_i)$
 $= \hat{a} \times x_i + \hat{b}$
- l'erreur observée, appelée "résidu" de la régression: $\hat{\epsilon}_i = y_i - \hat{y}_i$
- La somme (et donc la moyenne) des résidus est nulle dans une régression avec constante:

$$\begin{aligned} \sum_i \hat{\epsilon}_i &= \sum_i [y_i - (\hat{a}x_i + \hat{b})] \\ &= n\bar{y} - n\hat{a}\bar{x} - n\hat{b} \\ &= n\bar{y} - n\hat{a}\bar{x} - n \times (\bar{y} - \hat{a}\bar{x}) \\ &= 0 \end{aligned}$$

40 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

- L'objectif est de construire des estimateurs qui minimisent la somme des carrés des résidus: $SCR = \sum_i \hat{\epsilon}_i^2$

$$= \sum_i (y_i - \hat{y}_i)^2$$
- Prédiction parfaite $\rightarrow SCR = 0$.
- Mais dans d'autre cas, qu'est-ce qu'une bonne régression ?
- A partir de quelle valeur de SCR peut-on dire que la régression est mauvaise ?
- Comparer la SCR avec une valeur de référence ?

41 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

\rightarrow décomposer la variance de Y:

On appelle *somme des carrés totaux* (SCT) la quantité suivante :

$$SCT = \sum_i (y_i - \bar{y})^2$$

$$= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2 \sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

Dans la régression avec constante, et uniquement dans ce cas, on montre que

$$2 \sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

42 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

\rightarrow décomposer la variance de Y:

On obtient dès lors l'équation d'analyse de variance :

$$SCT = SCE + SCR$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

43 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

\rightarrow décomposer la variance de Y: *Interprétation Graphique:*

$SCT = \sum_i (Y_i - \bar{Y})^2$

$SCR = \sum_i (Y_i - \hat{Y}_i)^2$

$SCE = \sum_i (\hat{Y}_i - \bar{Y})^2$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

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

44 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

- Comment interpréter ces quantités ?
- ✓ **SCT** est la somme des carrés totaux. Elle indique la variabilité totale de Y , i.e. l'information disponible dans les données.
- ✓ **SCE** est la somme des carrés expliqués. Elle indique la variabilité expliquée par le modèle, i.e. la variation de Y expliquée par X .
- ✓ **SCR** est la somme des carrés résiduels. Elle indique la variabilité non-expliquée (résiduelle) par le modèle, i.e. l'écart entre les valeurs observées de Y et celles prédites par le modèle.

45 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

- Situations extrêmes:
- ✓ Le meilleur des cas: $SCR = 0 \rightarrow SCT = SCE$:
 - les variations de Y sont complètement expliquées par celles de X .
 - On a un modèle parfait
 - la droite de régression passe exactement par tous les points du nuage ($y_i^A = y_i$).
- ✓ Le pire des cas: $SCE = 0$:
 - X n'apporte aucune information sur Y .
 - $y_i^A = \bar{Y}$, la meilleure prédiction de Y est sa propre moyenne.

46 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Décomposition de la variance – Equation d'analyse de variance

- Tableau d'analyse de variance:

Source de variation	Somme des carrés
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$

Tableau simplifié d'analyse de variance

47 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction			
SPSS			

Coefficient de détermination

- Coefficient de détermination:

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$
- un indicateur synthétique calculé à partir de l'équation d'analyse de variance.
- Il indique la proportion de variance de Y expliquée par le modèle.
- R^2 proche de 1:
 - bon modèle,
 - la connaissance des valeurs de X permet de deviner avec précision celle de Y
- R^2 proche de 0:
 - X n'apporte pas d'informations utiles (intéressantes) sur Y
 - la connaissance des valeurs de X ne nous dit rien sur celles de Y .

48 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Coefficient de corrélation linéaire multiple

- Le coefficient de corrélation linéaire multiple est la racine carrée du coefficient de détermination: $R = \sqrt{R^2}$
- Dans le cas de la régression simple (et uniquement dans ce cas), R est le coefficient de corrélation r_{yx} entre Y et X. Son signe est déni par la pente a^\wedge de la régression: $r_{yx} = \text{signe}(a^\wedge) \times R$

49 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** Propriétés Inférence Intervalle de Prédiction SPSS

Exemple: Rendement Agricole

i	Y	X	Y*	epsion*	(Y-YB)*	(Y*-YB)*	(Y-Y*)^2
1	16	20	19.674	-2.674	102.010	55.148	7.149
2	18	24	21.530	-3.530	65.610	20.884	12.481
3	23	28	24.386	-1.386	9.610	2.937	1.922
4	24	22	20.102	3.898	4.410	35.977	15.195
5	28	32	27.242	0.758	3.610	1.305	0.574
6	29	28	24.386	4.614	8.410	2.937	21.286
7	26	32	27.242	-1.242	0.010	1.305	1.544
8	31	36	30.099	0.901	24.010	15.990	0.812
9	32	41	33.669	-1.669	34.810	57.289	2.785
10	34	41	33.669	0.331	62.410	57.289	0.110

Moyenne	26.1
a*	0.71405
b*	4.39277

314.900	251.061	63.839
SCT	SCE	SCR

R² = 0.797273
 Racine(R²) = 0.892901
 Corrély(x) = 0.892901

50 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** **Propriétés** Inférence Intervalle de Prédiction SPSS

Propriété des estimateurs

- Deux propriétés importantes pour l'évaluation d'un estimateur:

- Est-ce qu'il est sans biais ? i.e. est-ce qu'en moyenne nous obtenons la vraie valeur du paramètre ?
- Est-ce qu'il est convergent? i.e. à mesure que la taille de l'échantillon augmente, l'estimation devient de plus en plus précise ?

51 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité **Modèle** **Propriétés** Inférence Intervalle de Prédiction SPSS

Propriété des estimateurs: Biais

- On dit que ϑ^\wedge est un estimateur sans biais de ϑ si $E[\vartheta^\wedge] = \vartheta$.
- Comment procéder à cette vérification pour a^\wedge et b^\wedge ?

→ a^\wedge et b^\wedge sont sans biais, si et seulement si les deux hypothèses suivantes sont respectées :

H1 L'exogène X n'est pas stochastique (X est non aléatoire) ;

H2.a $E(e_i) = 0$, l'espérance de l'erreur est nulle.

52 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés

Propriété des estimateurs: variance - convergence

- Un estimateur $\hat{\theta}^A$ sans biais de θ est convergent si et seulement si $V(\hat{\theta}) \xrightarrow{n \rightarrow \infty} 0$
- La variance est $V(\hat{a}) = E[(\hat{a} - a)^2]$

→ \hat{a} et \hat{b} sont convergents, si et seulement si les deux hypothèses suivantes sont respectées :

H2.b $V(\epsilon_i) = \sigma^2$. C'est l'hypothèse d'homoscédasticité.

H2.d $COV(\epsilon_i, \epsilon_i') = E(\epsilon_i \epsilon_i') = 0$. C'est l'hypothèse de non-autocorrélation des erreurs.

53 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés

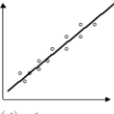
Propriété des estimateurs: remarques

- Estimateurs plus précis → les variances plus petites :
 - La variance de l'erreur est faible, i.e. la régression est de bonne qualité.
 - La dispersion des X est forte, i.e. les points recouvrent bien l'espace de représentation.
 - Le nombre d'observations n est élevé.

54 H. Benbrahim ENSIAS

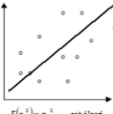
Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés

Propriété des estimateurs: remarques



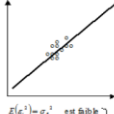
(1)

$E(\epsilon_i^2) = \sigma^2$ est faible
→ $V(\hat{\beta})$ est faible, modèle « stable »



(2)

$E(\epsilon_i^2) = \sigma^2$ est élevé
→ $V(\hat{\beta})$ est conséquemment élevée
Cette élévation est compensée par la valeur élevée de $\sum (x_i - \bar{x})^2$



(3)

$E(\epsilon_i^2) = \sigma^2$ est faible
 $\sum (x_i - \bar{x})^2$ est faible

55 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés

Théorème de Gauss - Markov

- Les estimateurs des Moindres carrés de la régression sont sans biais et convergents.
- Parmi les estimateurs linéaires sans biais de la régression, les estimateurs MC sont à variance minimale, i.e. il n'existe pas d'autres estimateurs linéaires sans biais présentant une plus petite variance.
- Les estimateurs des MC sont BLUE (best linear unbiased estimator).

56 H. Benbrahim ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Evaluation globale de la régression

- Evaluation de la qualité de l'ajustement:
 - la décomposition de la variance
 - le coefficient de détermination R^2 , i.e. dans quelle proportion la variabilité de Y pouvait être expliquée par X .
- Est-ce que la régression est globalement significative ?
 - Est-ce que X emmène **significativement** de l'information sur Y
 - Est-ce que R^2 est représentative d'une relation linéaire réelle dans la population, ou bien juste une simple fluctuation d'échantillonnage ?
- Ou bien: considérer le test d'évaluation globale comme un test de significativité de R^2 :
 - dans quelle mesure R^2 calculé sur un échantillon s'écarte réellement de la valeur 0 ?

57

H. Benbrahim

ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Tableau d'analyse de Variance - Test de significativité globale

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$CME = \frac{SCE}{1}$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n - 2$	$CMR = \frac{SCR}{n-2}$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$	$n - 1$	-

Tableau d'analyse de variance pour la régression simple

- Degrés De Liberté: le nombre de termes impliqués dans les sommes (le nombre d'observations) moins le nombre de paramètres estimés dans cette somme.
 - SCT: estimation de la moyenne $\bar{y} \rightarrow DDL = n-1$.
 - SCR: coefficients estimés \hat{a} et \hat{b} pour obtenir la projection $\hat{y}_i \rightarrow DDL = n-2$.
 - SCE: $SCT - SCR \rightarrow DDL = (n-1) - (n-2) = 1$.

58

H. Benbrahim

ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Test de significativité globale de la régression

- La statistique F:
$$F = \frac{CME}{CMR} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}$$
 - Cette statistique indique si la variance expliquée est significativement supérieure à la variance résiduelle.
 - l'explication emmenée par la régression traduit une relation qui existe réellement dans la population.
- La statistique F:
$$F = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{n-2}}$$

59

H. Benbrahim

ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Test de significativité globale de la régression

- Distribution sous H_0 :
 - SCE est distribué selon un $\chi^2(1)$
 - SCR est distribué selon un $\chi^2(n-2)$
 - $$F \equiv \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{n-2}} \equiv F(1, n-2)$$
 - Sous H_0 , F est donc distribué selon une loi de Fisher à $(1, n-2)$ degrés de liberté.
- La région critique du test:
 - correspondant au rejet de H_0
 - au risque α est définie pour les valeurs anormalement élevées de F
 $\rightarrow R.C. : F > F_{1-\alpha}(1, n-2)$
- Décision à partir de la p-value:
 - la probabilité critique (p-value) $\alpha' =$ probabilité que la loi de Fisher dépasse la statistique calculée F .
 - la règle de décision au risque α : $R.C. : \alpha' < \alpha$

60

H. Benbrahim

ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Exemple : les rendements agricoles

i	Y	X	Y*	spillon*	(Y-YIB)*	(Y*-YIB)*	(Y-Y*)²
1	16	20	18.674	-2.674	102.010	55.148	7.149
2	18	24	21.530	-3.530	65.010	20.884	12.461
3	23	26	24.386	-1.386	9.610	2.937	1.922
4	24	22	20.102	3.898	4.410	35.977	15.195
5	28	30	27.262	0.738	9.610	1.305	0.574
6	29	28	24.386	4.614	9.410	2.937	21.286
7	26	32	27.242	-1.242	0.010	1.305	1.544
8	31	36	30.099	0.901	24.010	15.990	0.812
9	32	41	33.669	-1.669	34.810	57.289	2.785
10	34	41	33.669	0.331	62.410	57.289	0.110

Moyenne 26.1

a* 0.71405
b* 4.39277

	SCY	SCX	SCN
Source	251.061	1	251.061
Expliquée	63.839	8	7.980
Réduite	187.222	1	243.081
Totale	314.900	9	

F 31.462

ddl 1

ddl 8

F 0.95 5.318

p-value 0.00050487

Conclusion : Le modèle est globalement significatif au risque 5%

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Exemple : les rendements agricoles

• Détail des calculs:

- ✓ $CME = \frac{SCE}{1} = \frac{251.061}{1} = 251.061$
- ✓ $CMR = \frac{SCR}{n-2} = \frac{63.839}{10-2} = 7.980$
- ✓ $F = \frac{CME}{CMR} = \frac{251.061}{7.980} = 31.462$

✓ nous comparons au quantile d'ordre $(1 - \alpha)$ de la loi $F(1, n - 2)$.
 ✓ Pour $\alpha = 5\%$, elle est égale à $F_{0.05}(1, 8) = 5.318$.
 → Nous concluons que le modèle est globalement significatif au risque 5%.
 → La relation linéaire entre Y et X est représentatif d'un phénomène existant réellement dans la population.

✓ la probabilité critique: $\alpha = 0.00050$, inférieure à $\alpha = 5\%$.
 → La conclusion est la même.

→ Il ne peut pas y avoir de contradictions entre ces deux visions.

62 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Distribution des paramètres estimés

• Pour étudier les coefficients estimés:

- calculer l'espérance et la variance des paramètres
- déterminer la loi de distribution.
- statistique inférentielle:
 - la définition des intervalles de variation à un niveau de confiance donné
 - la mise en place des tests d'hypothèses → les tests de significativité.

63 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Distribution de a^\wedge

• On a: $\hat{a} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ et $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$

• $y_i = ax_i + b + \varepsilon_i$ suit aussi une loi normale, et a^\wedge étant une combinaison linéaire des y_i

→ $\frac{\hat{a} - a}{\sigma_{\hat{a}}} \equiv \mathcal{N}(0, 1)$

• aussi: $\sigma_{\hat{a}}^2 = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

• mais $\sigma_\varepsilon^2 = ?$

→ Pour obtenir une estimation calculable sur un échantillon de données de l'écart-type $\hat{\sigma}_{\hat{a}}$ du coefficient a^\wedge → produire une estimation de l'écart type de l'erreur $\hat{\sigma}_\varepsilon$.
 La variance estimée: $\hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

64 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction SPSS			

Distribution de b^{\wedge}

- de même on a: $\frac{\hat{b} - b}{\sigma_b} \equiv \mathcal{N}(0, 1)$
- et: $\sigma_b^2 = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right]$

65 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction SPSS			

Estimateur sans biais de la variance de l'erreur

- on a le résidu: $\varepsilon_i = y_i - \hat{y}_i$
 $= ax_i + b + \varepsilon_i - (\hat{a}x_i + \hat{b})$
 $= \varepsilon_i - (\hat{a} - a)x_i - (\hat{b} - b)$
- on montre que: $E \left[\sum_i \varepsilon_i^2 \right] = (n-2)\sigma_\varepsilon^2$
- Estimateur sans biais: $\hat{\sigma}_\varepsilon^2 = \frac{\sum_i \varepsilon_i^2}{n-2} = \frac{SCR}{n-2}$

66 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction SPSS			

Distribution de la variance de l'erreur

- On a par hypothèse: $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$
- $\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \equiv \mathcal{N}(0, 1)$
- $\left(\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \right)^2 = \chi^2(1)$
- $\sum_i \left(\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \right)^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-2)$
- $\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \frac{\chi^2(n-2)}{n-2}$

67 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
Généralité	Modèle	Propriétés	Inférence
Intervalle de Prédiction SPSS			

Distribution de a^{\wedge} et b^{\wedge} :

- On a: $\frac{\hat{\sigma}_a^2}{\sigma_a^2} \equiv \frac{\chi^2(n-2)}{n-2}$
- $\frac{\hat{\sigma}_a^2}{\sigma_a^2} = \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \frac{\chi^2(n-2)}{n-2}$
- $\frac{\hat{a} - a}{\hat{\sigma}_a} \equiv \mathcal{T}(n-2)$, de même: $\frac{\hat{b} - b}{\hat{\sigma}_b} \equiv \mathcal{T}(n-2)$
- loi de Student est définie par un rapport entre une loi normale et la racine carrée d'une loi du χ^2 normalisée par ses degrés de liberté.
- $\frac{\frac{\hat{a}-a}{\hat{\sigma}_a}}{\frac{\frac{\hat{\sigma}_\varepsilon}{\sigma_\varepsilon}}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}} \equiv \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}$
- $\frac{\hat{a} - a}{\hat{\sigma}_a} \equiv \mathcal{T}(n-2)$

68 H. Benbrahim ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Test de significativité de a:

- Le test de significativité de la pente a consiste à vérifier l'influence réelle de l'exogène X sur l'endogène Y .
- Les hypothèses à confronter :
$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$$
- Statistique de test: $t_a = \frac{\hat{a}}{\hat{\sigma}_a}$
suit une loi de Student à $(n - 2)$ degrés de liberté.
- La région critique (de rejet de H_0) au risque α s'écrit: $R.C. : |t_a| > t_{1-\frac{\alpha}{2}}$
Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi de Student. Il s'agit d'un test bilatéral.

69

H. Benbrahim

ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Exemple: Rendement Agricole:

i	Y	X	Y^	epsilon^	(epsilon^)^2	(X.XB)^2
1	16	20	18.674	-2.674	7.149	108.16
2	18	24	21.530	-3.530	12.461	40.96
3	23	28	24.386	-1.386	1.922	5.76
4	24	22	20.102	3.898	15.195	70.56
5	28	32	27.242	0.758	0.574	2.56
6	29	28	24.386	4.614	21.295	5.76
7	26	32	27.242	-1.242	1.544	2.56
8	31	36	30.099	0.901	0.812	31.36
9	32	41	33.669	-1.669	2.785	112.36
10	34	41	33.669	0.331	0.110	112.36
Somme						492.4
Moyenne	26.1	30.4	SCR		63.839	
a^	0.71405		[(sigma^)^2 eps		7.980	
b^	4.39277		sigma^*(eps)		2.825	

70

H. Benbrahim

ENSIAS

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Exemple: Rendement Agricole:

- l'estimation de la variance de l'erreur:
$$\hat{\sigma}_\epsilon^2 = \frac{SCR}{n - 2} = \frac{63.839}{8} = 7.980 \quad \hat{\sigma}_\epsilon = \sqrt{7.980} = 2.825$$
- $$\hat{\sigma}_a = \sqrt{\frac{\hat{\sigma}_\epsilon^2}{\sum_i (x_i - \bar{x})^2}}$$
$$= \sqrt{\frac{7.980}{492.4}} = \sqrt{0.01621} = 0.12730$$
$$t_a = \frac{\hat{a}}{\hat{\sigma}_a} = \frac{0.71405}{0.12730} = 5.60909$$

Au risque $\alpha = 5\%$, le seul critique pour la loi de Student à $(n - 2)$ degrés de liberté pour un test bilatéral⁶ est $t_{1-\frac{\alpha}{2}} = 2.30600$. Puisque $|5.60909| > 2.30600$, nous concluons que la pente est significativement non nulle au risque 5%.

Définition

Régression Linéaire Simple

Régression Linéaire Multiple

Conclusion

Généralité

Modèle

Propriétés

Inférence

Intervalle de Prédiction

SPSS

Exemple: Rendement Agricole:

ESTIMATION

a	0.714053819
b	4.392770106

sigma(epsilon)

7.979843823

sigma(a^)	0.016200019	sigma(a^)	0.127302862
sigma(b^)	15.77493893	sigma(b^)	3.971767696

d.f.

8

t theorique (bilatéral a 5%)

2.306004133

t(a^)	5.609093169	rejet H0
t(b^)	1.10599875	

$t_a = \frac{\hat{a}}{\hat{\sigma}_a} = \frac{0.714}{0.127} = 5.609$

$t_{1-\frac{\alpha}{2}}(8) = t_{1-0.05/2}(8) = t_{0.975}(8) = 2.306$

Puisque

$|t_a| > t_{1-\frac{\alpha}{2}}$

Rejet de $H_0 : a = 0$

72

H. Benbrahim

ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Intervalle de confiance de la droite de la régression

- Les coefficients formant le modèle sont entachés d'incertitude
→ la droite de régression l'est également.
- L'objectif est de produire un intervalle de confiance de la droite de régression.

↔ au calcul de l'intervalle de confiance de la prédiction de la moyenne de Y conditionnellement X .

- ✓ C'est l'intervalle de confiance de ce que l'on a modélisé avec la droite
- ✓ à ne pas confondre avec l'intervalle de confiance d'une prédiction lorsque l'on fournit la valeur x_i pour un nouvel individu i n'appartenant pas à l'échantillon.
- L'intervalle de confiance au niveau $(1-\alpha)$ de la droite de régression:

$$\hat{a} \times x_i + \hat{b} \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_\epsilon \sqrt{\frac{1}{n} + \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}}$$

73 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Exemple: Rendement Agricole:

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
n = 10	34	41

Y ²	epsilon ²	(epsilon) ²
18 674	-2 674	7 149
21 530	-3 530	12 461
24 386	-1 386	1 922
20 102	3 898	15 195
27 242	0 758	0 574
24 386	4 614	21 286
27 242	-1 242	1 544
30 099	0 901	0 812
33 669	-1 669	2 785
33 669	0 331	0 110

(X-XB) * 2	b basse	b haute
108 16	14 99	22 36
40 96	18 74	24 32
5 76	22 21	26 56
70 56	16 89	23 32
2 56	25 13	29 36
5 76	22 21	26 56
2 56	25 13	29 36
31 36	27 46	32 73
112 36	29 94	37 40
112 36	29 94	37 40

Moyenne 30.4

SCR 63 8387

sigma(epsilon) 2 8249

Somme 492.4

n 10

a⁰ 0.71405

b⁰ 4.39277

1.0 975(0) 2 30600

74 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Exemple: Rendement Agricole:

$$b.b.(p_{Y/X=x_i}) = 18.674 - 2.30600 \times 2.8249 \times \sqrt{\frac{1}{10} + \frac{(20-30.4)^2}{492.4}} = 14.99$$

$$b.h.(p_{Y/X=x_i}) = 18.674 + 2.30600 \times 2.8249 \times \sqrt{\frac{1}{10} + \frac{(20-30.4)^2}{492.4}} = 22.36$$

- Il y a 95% de chances que la droite soit comprise entre les deux courbes bleues.
- La droite ne peut être placée n'importe où dans la zone délimitée.
- elle pivote forcément autour du barycentre.

75 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés **Inférence** Intervalle de Prédiction SPSS

Prédiction et Intervalle de Prédiction

- Régression?
 - analyse structurelle
 - interprétation des coefficient
 - utilisée pour la prédiction ou prévision:
 - ✓ Pour un nouvel individu donné, à partir de la valeur de l'exogène X , connaître la valeur que prendrait l'endogène Y .
- ✓ Pour un nouvel individu i , qui n'appartient pas à l'échantillon de données ayant participé à l'élaboration du modèle, connaissant la valeur de x_i , on cherche à obtenir la prédiction ay_i .

76 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés
		Inférence	Intervalle de Prédiction
			SPSS

Prédiction ponctuelle

- Pour prédire Y à partir d'une valeur connue X:
- On applique directement l'équation de régression: $\hat{y}_{i*} = \hat{y}(x_{i*})$

$$= \hat{a} \times x_{i*} + \hat{b}$$
- La prédiction est sans biais: $E[\hat{y}_{i*}] = y_{i*}$.

En effet,

$$\begin{aligned} \hat{\epsilon}_{i*} &= \hat{y}_{i*} - y_{i*} \\ &= \hat{a}x_{i*} + \hat{b} - (ax_{i*} + b + \epsilon_{i*}) \\ &= (\hat{a} - a)x_{i*} + (\hat{b} - b) - \epsilon_{i*} \end{aligned}$$

$$E(\hat{\epsilon}_{i*}) = E[(\hat{a} - a)x_{i*} + (\hat{b} - b) - \epsilon_{i*}]$$

$$= x_{i*}E(\hat{a} - a) + E(\hat{b} - b) - E(\epsilon_{i*})$$

Les EMC sont sans biais L'erreur du modèle est nulle par hypothèse

77 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés
		Inférence	Intervalle de Prédiction
			SPSS

Prédiction par intervalle

- Prédiction ponctuelle est intéressante, mais avec quel degré de confiance?
- une intervalle de prédiction en lui associant une probabilité de recouvrir la vraie valeur y_{i*} .
- Connaître
 - la variance de l'erreur de prédiction
 - la loi de distribution de l'erreur.

78 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés
		Inférence	Intervalle de Prédiction
			SPSS

Prédiction par intervalle – Variance de l'erreur de prédiction

Variance de l'erreur de prévision

Puisque $\hat{\epsilon}_{i*} = \hat{y}_{i*} - y_{i*}$
 $E(\hat{\epsilon}_{i*}) = 0$

On montre

$$V(\hat{\epsilon}_{i*}) = E(\hat{\epsilon}_{i*}^2) = \sigma_{\epsilon}^2 \left[1 + \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right] = \sigma_{\epsilon}^2$$

D'où la variance estimée de l'erreur de prévision

$$\hat{\sigma}_{\hat{\epsilon}_{i*}}^2 = \hat{\sigma}_{\epsilon}^2 \left[1 + \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$

Remarque :

$$h_{i*} = \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

est le LEVIER de l'observation i^*
 (Il joue un rôle très important dans la régression : points atypiques).

79 H. Benbrahim ENSIAS

Définition	Régression Linéaire Simple	Régression Linéaire Multiple	Conclusion
	Généralité	Modèle	Propriétés
		Inférence	Intervalle de Prédiction
			SPSS

Prédiction par intervalle – Variance de l'erreur de prédiction

La variance de l'erreur sera d'autant plus faible que :

- (1) $\hat{\sigma}_{\epsilon}^2 = \frac{SCR}{n-2}$ est petit c.-à-d. la droite ajuste bien le nuage de points.
- (2) $(x_{i*} - \bar{x})^2$ est petit c.-à-d. le point est proche du centre de gravité du nuage.
- (3) $\sum_i (x_i - \bar{x})^2$ est grand c.-à-d. la dispersion des points est grande.
- (4) n est grand c.-à-d. le nombre d'observations ayant servi à la construction du modèle est élevé.

80 H. Benbrahim ENSIAS

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés Inférence **Intervalle de Prédiction** SPSS

Prédiction par intervalle – la distribution de la variance de l'erreur de prédiction

Puisque $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ $\Rightarrow \hat{\varepsilon}_p = \hat{y}_p - y_p \sim N(0, \sigma_\varepsilon^2 \sqrt{1 + h_p})$

$\Rightarrow (n-2) \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \sim \chi^2(n-2)$

$\Rightarrow \frac{\hat{y}_p - y_p}{\hat{\sigma}_{\hat{y}_p}} \sim \mathcal{N}(0, 1)$ Rapport d'une loi normale avec un KHI-2 normalisé

$\Rightarrow \hat{y}_{i^*} \pm t_{1-\alpha/2} \times \hat{\sigma}_{\hat{y}_p}$ Intervalle de confiance au niveau $(1-\alpha)$

Définition **Régression Linéaire Simple** Régression Linéaire Multiple Conclusion

Généralité Modèle Propriétés Inférence **Intervalle de Prédiction** SPSS

Exemple: Rendement Agricole

Rendements agricoles – $x^* = 38$

Prédiction ponctuelle $\rightarrow \hat{y}_p = dx_p + b$
 $= 0.714 \times 38 + 4.39$
 $= 31.5268$

Y	X	Y-X	(Y-X) ²	Y(X-Y)	Y(X-Y) ²	Y(X-Y) ³
20	20	0	0	0	0	0
22	22	0	0	0	0	0
24	22	2	4	44	88	176
26	22	4	16	88	352	704
28	22	6	36	132	792	1584
30	22	8	64	176	1408	2816
32	22	10	100	220	2200	4400
34	22	12	144	264	3168	6336
36	22	14	196	308	4312	8848
38	22	16	256	352	5632	12288
40	22	18	324	396	7128	17712
42	22	20	400	440	8800	25200
44	22	22	484	484	10648	33472
46	22	24	576	528	12672	42912
48	22	26	676	572	14976	54656
50	22	28	784	616	17568	68800
52	22	30	900	660	19440	85200
54	22	32	1024	704	21632	103936
56	22	34	1156	748	24160	125056
58	22	36	1296	792	27072	148608
60	22	38	1444	836	30368	175744
62	22	40	1600	880	34080	206400
64	22	42	1764	924	38208	241728
66	22	44	1936	968	42752	281856
68	22	46	2116	1012	47712	326800
70	22	48	2304	1056	53088	376608
72	22	50	2500	1100	58880	431400
74	22	52	2704	1144	65088	491328
76	22	54	2916	1188	71712	556512
78	22	56	3136	1232	78752	627008
80	22	58	3364	1276	86208	702864
82	22	60	3600	1320	94080	784128
84	22	62	3844	1364	102368	870848
86	22	64	4096	1408	111088	964064
88	22	66	4356	1452	120240	1063824
90	22	68	4624	1496	129824	1170176
92	22	70	4900	1540	139840	1283264
94	22	72	5184	1584	150288	1403136
96	22	74	5476	1628	161168	1529824
98	22	76	5776	1672	172480	1663264
100	22	78	6084	1716	184224	1803504
102	22	80	6400	1760	196400	1950576
104	22	82	6724	1804	209008	2104512
106	22	84	7056	1848	222048	2265344
108	22	86	7396	1892	235520	2433104
110	22	88	7744	1936	249440	2607824
112	22	90	8100	1980	263808	2789520
114	22	92	8464	2024	278624	2978224
116	22	94	8836	2068	293888	3174064
118	22	96	9216	2112	309600	3377072
120	22	98	9604	2156	325760	3587280
122	22	100	10000	2200	342368	3804720
124	22	102	10404	2244	359424	4029424
126	22	104	10816	2288	376944	4261424
128	22	106	11236	2332	394928	4499760
130	22	108	11664	2376	413376	4744464
132	22	110	12100	2420	431288	4995568
134	22	112	12544	2464	449664	5253008
136	22	114	12996	2508	468504	5516832
138	22	116	13456	2552	487808	5787088
140	22	118	13924	2596	507576	6063808
142	22	120	14400	2640	527808	6346928
144	22	122	14884	2684	548496	6636384
146	22	124	15376	2728	569648	6932224
148	22	126	15876	2772	591168	7234496
150	22	128	16384	2816	613056	7543136
152	22	130	16896	2860	635312	7858192
154	22	132	17416	2904	657936	8179696
156	22	134	17944	2948	680928	8507680
158	22	136	18480	2992	704288	8842176
160	22	138	19024	3036	728016	9183216
162	22	140	19576	3080	752112	9530832
164	22	142	20136	3124	776576	9885072
166	22	144	20704	3168	801408	10245968
168	22	146	21280	3212	826608	10613536
170	22	148	21864	3256	852176	10987808
172	22	150	22456	3300	878112	11368800
174	22	152	23056	3344	904416	11756544
176	22	154	23664	3388	931088	12150976
178	22	156	24280	3432	958128	12552128
180	22	158	24904	3476	985536	12959936
182	22	160	25536	3520	1013312	13374432
184	22	162	26176	3564	1041456	13795648
186	22	164	26824	3608	1069968	14223616
188	22	166	27480	3652	1098848	14658288
190	22	168	28144	3696	1128096	15109696
192	22	170	28816	3740	1157712	15567872
194	22	172	29496	3784	1187704	16032848
196	22	174	30184	3828	1218072	16504656
198	22	176	30880	3872	1248816	16983328
200	22	178	31584	3916	1279936	17468896
202	22	180	32296	3960	1311440	17961392
204	22	182	33016	4004	1343328	18460848
206	22	184	33744	4048	1375600	18967304
208	22	186	34480	4092	1408256	19480800
210	22	188	35224	4136	1441296	19999376
212	22	190	35976	4180	1474720	20524064
214	22	192	36736	4224	1508528	21054912
216	22	194	37504	4268	1542720	21591952
218	22	196	38280	4312	1577296	22135216
220	22	198	39064	4356	1612256	22684736
222	22	200	39856	4400	1647600	23240544
224	22	202	40656	4444	1683328	23802688
226	22	204	41464	4488	1719440	24371184
228	22	206	42280	4532	1755936	24946064
230	22	208	43104	4576	1792816	25527360
232	22	210	43936	4620	1830080	26115008
234	22	212	44776	4664	1867728	26709056
236	22	214	45624	4708	1905760	27309536
238	22	216	46480	4752	1944176	27916496
240	22	218	47344	4796	1982976	28529984
242	22	220	48216	4840	2022160	29149936
244	22	222	49096	4884	2061728	29776304
246	22	224	49984	4928	2101680	30409136
248	22	226	50880	4972	2142016	31048464
250	22	228	51784	5016	2182736	31694320
252	22	230	52696	5060	2223840	32346736
254	22	232	53616	5104	2265328	33005744
256	22	234	54544	5148	2307200	33671296
258	22	236	55480	5192	2349456	34343424
260	22	238	56424	5236	2392096	35022064
262	22	240	57376	5280	2435120	35707248
264	22	242	58336	5324	2478528	36398912
266	22	244	59304	5368	2522320	37097088
268	22	246	60280	5412	2566496	37801808
270	22	248	61264	5456	2611056	38513104
272	22	250	62256	5500	2656000	39230912
274	22	252	63256	5544	2701328	39955264
276	22	254	64264	5588	2747040	40686208
278	22	256	65280	5632	2793136	41423680
280	22	258	66304	5676	2839616	42167712
282	22	260	67336	5720	2886480	42918336
284	22	262	68376	5764	2933728	43675584
286	22	264	69424	5808	2981360	44439408
288	22	266	70480	5852	3029376	45209744
290	22	268	71544	5896	3077776	45986624
292	22	270	72616	5940	3126560	46769984
294	22	272	73696	5984	3175728	47559856
296	22	274	74784	6028	3225280	48356272
298	22	276	75880	6072	3275216	49159264
300	22	278	76984	6116	3325536	49968768
302	22	280	78096	6160	3376240	50784816
304	22	282	79216	6204	3427328	51607456
306	22	284	80344	6248	3478800	52436624
308	22	286	81480	6292	3530656	53272368
310	22	288	82624	6336	3582896	54114720
312	22	290	83776	6380	3635520	54962720
314	22	292	84936	6424	3688528	55816400
316	22	294	86104	6468	3741920	56675776
318	22	296	87280	6512	3795696	57540880
320	22	298	88464	6556	3849856	58412736
322	22	300	89656	6600	3904400	59291280
324	22	302	90856	6644	3959328	60176544
326	22	304	92064	6688	4014640	61068560
328	22	306	93280	6732	4070336	61967360
330	22	308	94504	6776	4126416	62872976
332	22	310	95736	6820	4182880	63785440
334	22	312	96976	6864	4239728	64704688
336	22	314	98224	6908	4296960	65630752
338	22	316	99480	6952	4354576	66563664
340	22	318	100744	6996	4412576	67503456
342	22	320	102016	7040	4470960	68450064
344	22	322	103296	7084	4529728	69403520
346	22	324	104584	7128	4588880	70363856
348	22	326	105880	7172	4648416	71331104
350	22	328	107184	7216	4708336	72305296
352	22	330	108496	7260	4768640	73286464
354	22	332	109816	7304	4829328	74274640
356	22	334	111144	7348	4890400	75269856</