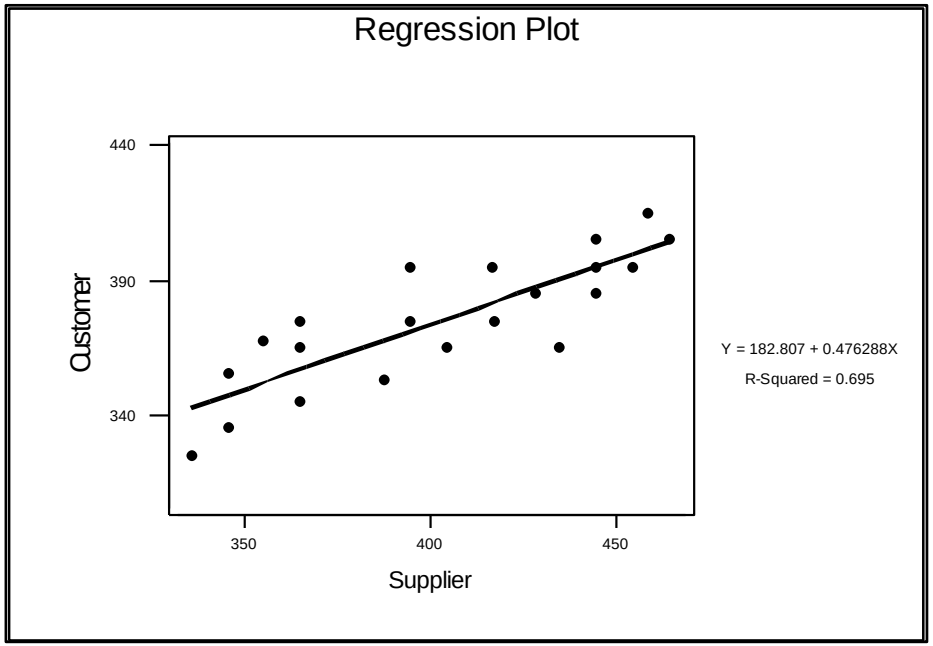
Corrélation & Régression simple



Terminologie

- **Corrélation: La corrélation est une mesure de la force d'association entre deux variables quantitatives (ex: pression et rendement). Elle mesure le degré de linéarité entre deux variables supposées complètement indépendantes l'une de l'autre.**

- **Le coefficient de corrélation, r, est toujours compris entre -1 et +1.**

  - Corrélation Positive : $0 < r < 1$; x et y les deux croisent ou décroisent ensemble.
  - Corrélation Négative: $-1 < r < 0$; x croisent et y décroisent ou vis versa.

- **Une valeur r de –1 indique une relation négative parfait, +1 indique une relation positive parfait, 0 indique l'absence de relation.**

Corrélation

La *covariance* est définie

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

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}.$$

Le *coefficient de corrélation* est la covariance divisée par les deux écart-types marginaux

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

$$-1 \leq r_{xy} \leq 1.$$

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Nombre échantillon & test de corrélation

En déterminant la taille de l' échantillon, toute corrélation supérieure à la valeur indiquée dans le tableau est considérée comme "importante" or ayant une signification statistique.

La valeur de p pour le coefficient de corrélation de Pearson utilise la loi de distribution t..

$H_o: \rho = 0$
$H_a: \rho \neq 0$

$$r_\alpha = \sqrt{\frac{t_\alpha^2}{n-2+t_\alpha^2}}$$

or

$$t_\alpha = \frac{\sqrt{n-2}\cdot r}{\sqrt{1-r^2}}$$

r coefficient de corrélation

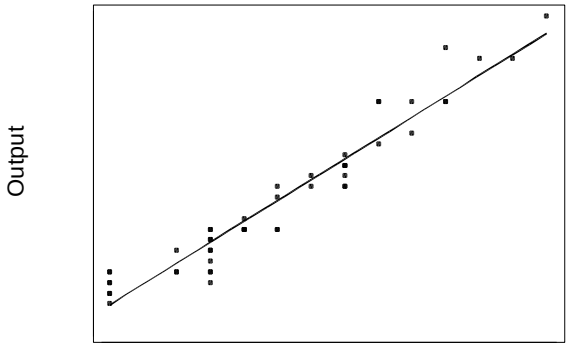
n Nombre d'observations

Sample Size	d.f.	Significance Level			
n	n-2	0.05	0.025	0.01	0.005
3	1	0.9877	0.9969	0.9995	0.9999
4	2	0.9000	0.9500	0.9800	0.9900
5	3	0.8054	0.8783	0.9343	0.9587
6	4	0.7293	0.8114	0.8822	0.9172
7	5	0.6694	0.7545	0.8329	0.8745
8	6	0.6215	0.7067	0.7887	0.8343
9	7	0.5822	0.6664	0.7498	0.7977
10	8	0.5494	0.6319	0.7155	0.7646
11	9	0.5214	0.6021	0.6851	0.7348
12	10	0.4973	0.5760	0.6581	0.7079
13	11	0.4762	0.5529	0.6339	0.6835
14	12	0.4575	0.5324	0.6120	0.6614
15	13	0.4409	0.5140	0.5923	0.6411
16	14	0.4259	0.4973	0.5742	0.6226
17	15	0.4124	0.4821	0.5577	0.6055
18	16	0.4000	0.4683	0.5425	0.5897
19	17	0.3887	0.4555	0.5285	0.5751
20	18	0.3783	0.4438	0.5155	0.5614
21	19	0.3687	0.4329	0.5034	0.5487
22	20	0.3598	0.4227	0.4921	0.5368
27	25	0.3233	0.3809	0.4451	0.4869
32	30	0.2960	0.3494	0.4093	0.4487
37	35	0.2746	0.3246	0.3810	0.4182
42	40	0.2573	0.3044	0.3578	0.3932
47	45	0.2429	0.2876	0.3384	0.3721
52	50	0.2306	0.2732	0.3218	0.3542
62	60	0.2108	0.2500	0.2948	0.3248
72	70	0.1954	0.2319	0.2737	0.3017
82	80	0.1829	0.2172	0.2565	0.2830
92	90	0.1726	0.2050	0.2422	0.2673
102	100	0.1638	0.1946	0.2301	0.2540

Coefficient de Correlation

Forte

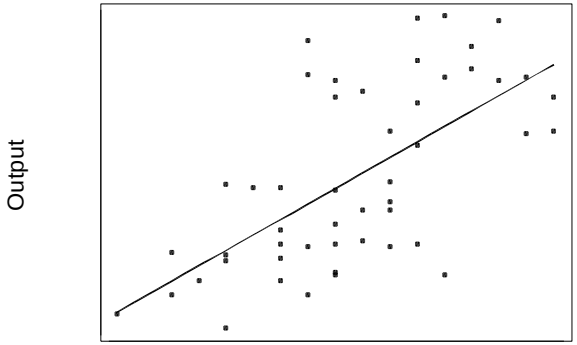
Strong Positive Correlation



r = 0.963

Modérée

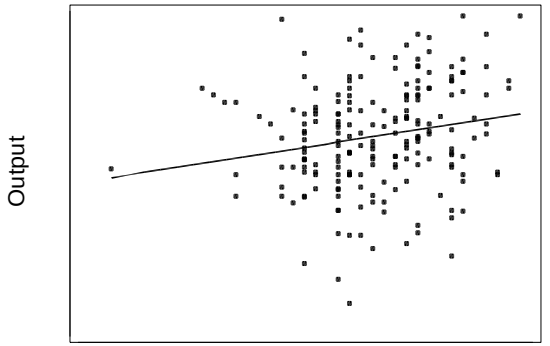
Moderate Postive Correlation



r = 0.646

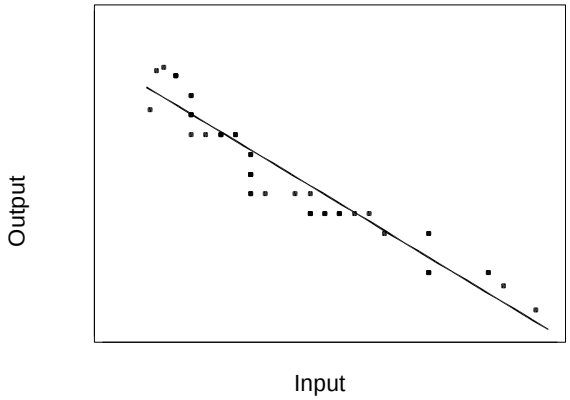
faible

Weak Positive Correlation



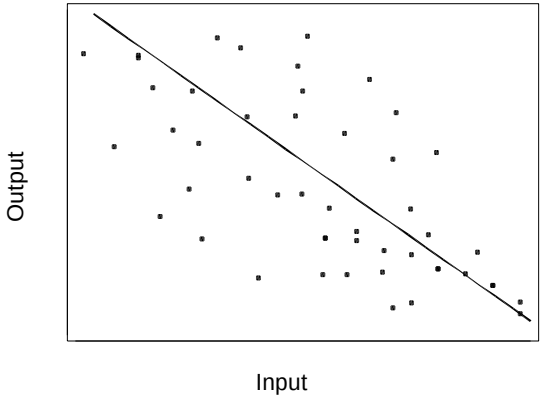
r = 0.196

Strong Negative Correlation



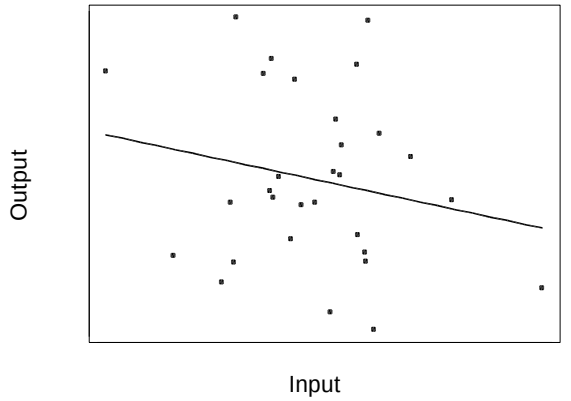
r = - 0.963

Moderate Negative Correlation



r = - 0.646

Weak Negative Correlation



r = - 0.196

Attention à déclarer la causalité

- Si nous établissons une causalité entre y et x, cela ne veut pas forcément dire que la variation de x a provoqué la variation de y.
- Une troisième variable peut « rôder » dans les parages et faire varier à la fois x et y.
- Exemple une forte corrélation entre la pression (x1) et le rendement d'un réacteur. Une forte corrélation négative entre la pression et le rendement a été établie. Toutefois:
  - Il existe une impureté (x2) qui n'est pas mesurée et varie d'un lot à l'autre
    - L'impureté provoque de l'écume, ce qui réduit le rendement
    - On augmente la pression pour réduire l'écume
    - La pression est une réaction à l'écume et n'a rien à voir avec le rendement.
- Y-a-t-il une corrélation entre la hauteur de l'herbe et la longueur des cheveux ? Devons-nous arroser les deux ?

- **La corrélation est un outil très utile dans les industries de transformation**
- **La corrélation est une mesure de la relation linéaire entre deux variables quantitatives**
- **Attention à ne pas toujours assumer la causalité**
- **La corrélation prépare aux techniques de régression**

Exercice

Un ingénieur souhaite étudier la relation entre le réglage spécifique d'une machine et la quantité d'énergie que la machine consomme, afin de déterminer un paramètre de fonctionnement optimum. L'ingénieur collecte des données relatives au réglage de la machine et à la consommation d'énergie durant le processus de fabrication.

Réglage machine	Conso énergie
11,15	21,6
13,3	18,5
14,2	17,2
15,7	18
18,9	15,7
19,4	15
21,4	13,8
21,7	14
23,5	12,4
24,3	102
25,3	11
26,4	9,2
26,7	7,6
27,9	6,9
29,1	4,5

1. **Y a-t-il une corrélation entre le nouveau réglage de la machine et la consommation d'énergie?**
2. **Etablir la régression entre réglage machine et consommation d'énergie**

	Réglage machine	Conso énergie	Xi-Xbar	Yi-Ybar	(Xi-Xbar)*(Yi-Ybar)	(Xi-Xbar)²	(Yi-Ybar)²	Regression= Yr	(Yr-Ybar)²	(Yi-Yr)^2	V résiduelle Yi-Yr
	11,15	21,6	- 10,11	8,56	- 86,57	102,28	73,27	21,44	70,50	0,03	0,16
	13,3	18,5	- 7,96	5,46	- 43,48	63,41	29,81	19,65	43,71	1,33	- 1,15
	14,2	17,2	- 7,06	4,16	- 29,38	49,89	17,31	18,90	34,39	2,90	- 1,70
	15,7	18	- 5,56	4,96	- 27,59	30,95	24,60	17,66	21,33	0,12	0,34
	18,9	15,7	- 2,36	2,66	- 6,29	5,59	7,08	15,00	3,85	0,49	0,70
	19,4	15	- 1,86	1,96	- 3,65	3,47	3,84	14,59	2,39	0,17	0,41
	21,4	13,8	0,14	0,76	0,10	0,02	0,58	12,93	0,01	0,76	0,87
	21,7	14	0,44	0,96	0,42	0,19	0,92	12,68	0,13	1,75	1,32
	23,5	12,4	2,24	- 0,64	- 1,43	5,00	0,41	11,18	3,45	1,48	1,22
	24,3	10,2	3,04	- 2,84	- 8,62	9,22	8,07	10,52	6,36	0,10	- 0,32
	25,3	11	4,04	- 2,04	- 8,23	16,29	4,16	9,69	11,23	1,72	1,31
	26,4	9,2	5,14	- 3,84	- 19,72	26,39	14,75	8,78	18,19	0,18	0,42
	26,7	7,6	5,44	- 5,44	- 29,58	29,56	29,59	8,53	20,37	0,86	- 0,93
	27,9	6,9	6,64	- 6,14	- 40,75	44,05	37,70	7,53	30,36	0,40	- 0,63
	29,1	4,5	7,84	- 8,54	- 66,93	61,41	72,93	6,53	42,33	4,14	- 2,03
Moyenne	**21,26**	**13,04**		**Somme**	**- 371,71**	**447,72**	**325,02**	**Moyenne**	**13,04**	**308,6**	**16,42**
cart type	**5,46**	**4,65**						**Ecart type**	**4,54**		

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = 371{,}71$$

r = -371,71 / 381,47 = - 0,974

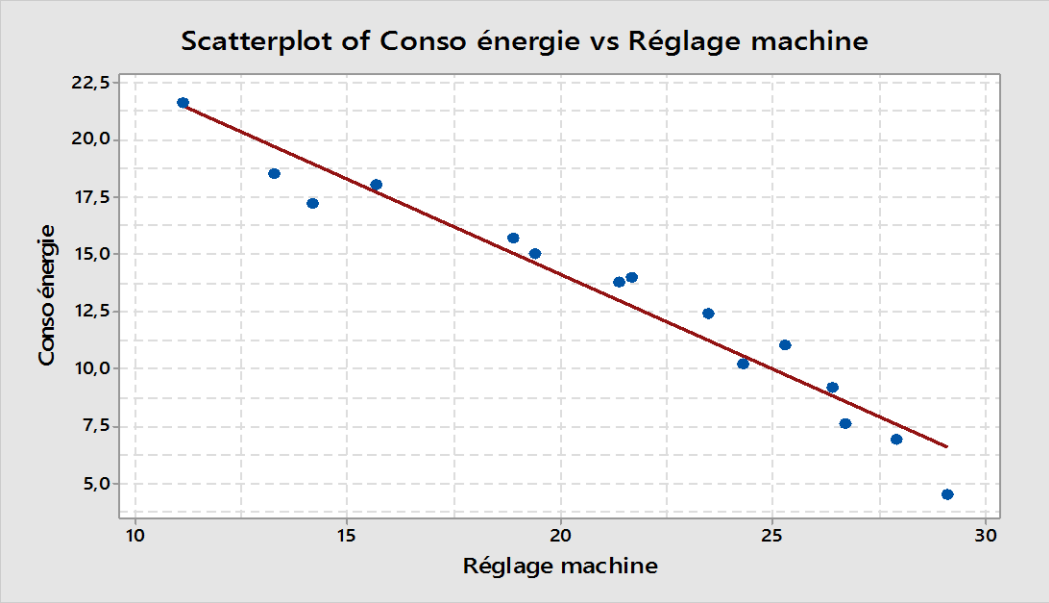
$$\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} = 381{,}47$$

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$T_{calculé}$ = 15,63

$T_{critique}$ = 2,145

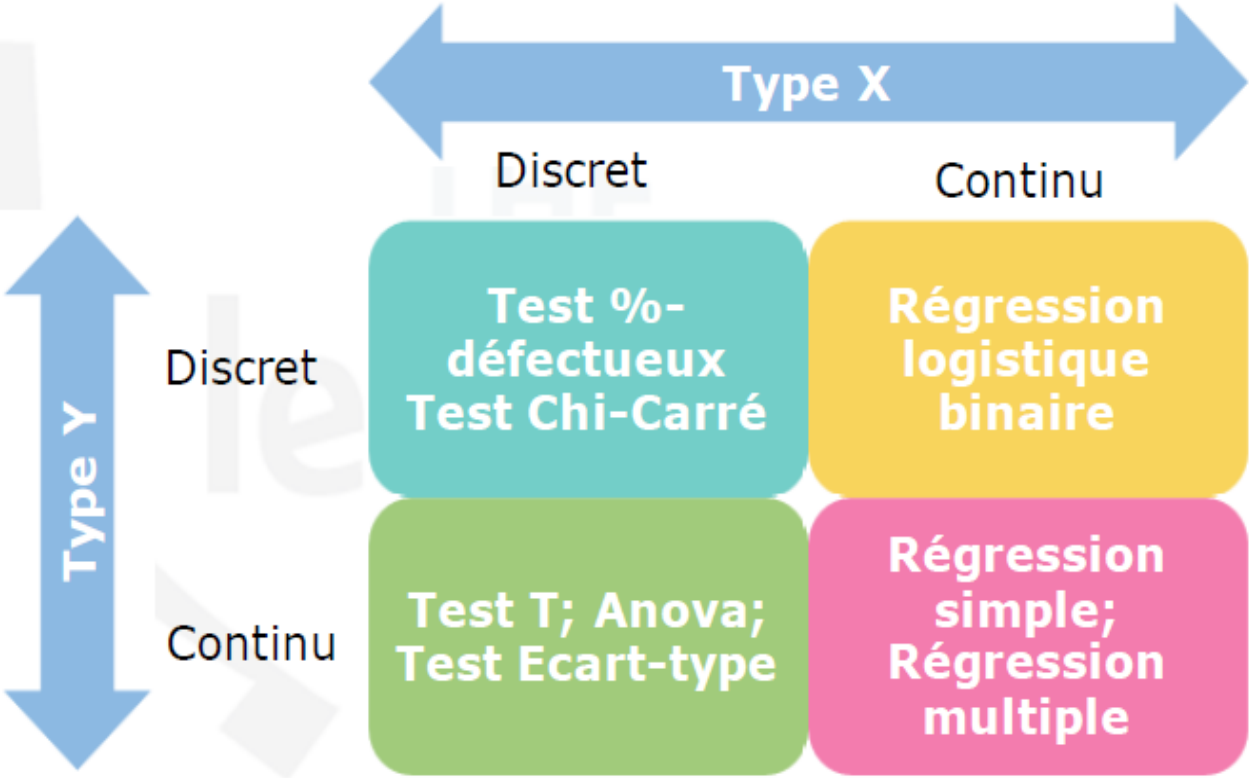
P= 0,000



Une corrélation forte et négative entre la consommation d'énergie et le nouveaux réglage de la machine

Régression simple

Tests d'hypothèse et type de données



Equation régression: **L'équation approprié, pas nécessairement linéaire qui permet de prédire les outputs en connaissant l'inputs**

- **L'analyse de régression est une méthode statistique pour investiguer et modéliser une relation entre variables.**

Supposons $Y = f(x_1, x_2, \ldots x_n)$

Cela décrit la relation exacte entre les variables d'entrée X's du processus (variables indépendantes) et la variable de sortie Y (variable dépendante):

- **L'analyse de Régression est utilisée pour faire une approximation de la relation, basée sur les données.**
- **Le Modèle de Régression est construit pour faire l'approximation exacte de la relation.**

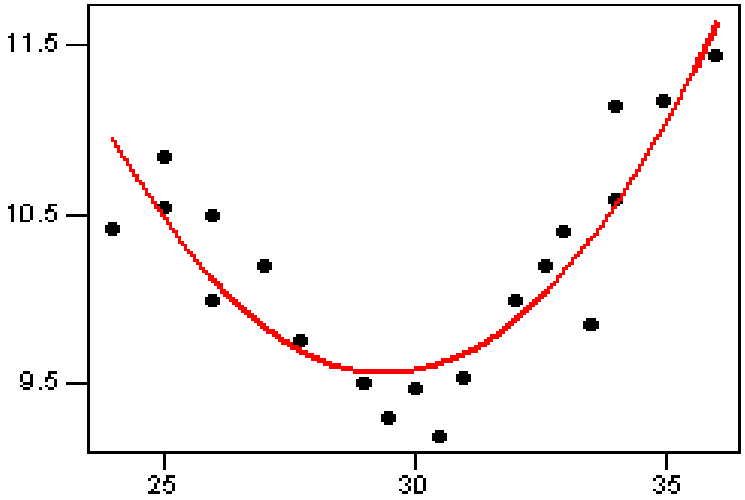
L'objectif majeur est de produire des prédictions précises et exactes de Y pour des Xs donnés.

Régression simple

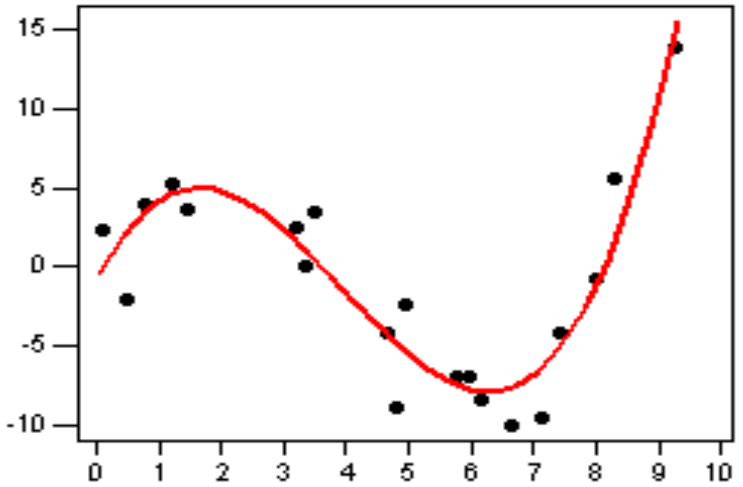
Tandis que la corrélation nous dit quelle association linéaire il y a entre deux variables, la régression définit plus précisément cette association.
La régression résulte en une équation qui utilise une ou plusieurs variable(s) pour expliquer la variation d'une autre variable.

Exemples de prédiction de l'équations:

- $Y = a + bx$ (Modèle linéaire)
- $Y = a + bx + cx^2$ (quadratique)
- $Y = a + bx + cx^2 + dx^3$ (cubique)



Quadratique



Cubique

Régression simple : Analyse numérique

La *droite de régression* est la droite qui ajuste au mieux un nuage de points au sens des moindres carrés.

$$y = a + bx$$

Les coefficients a et b qui minimisent le critère des moindres carrés sont donnés par :

$$b = \frac{s_{xy}}{s_x^2} \quad et \quad a = \bar{y} - b\bar{x}$$

Les résidus de la régression définis par

$$e_i = y_i - a - bx_i.$$

Les *valeurs ajustées* sont obtenues au moyen de la droite de régression :

$$e_i = y_i - y_i^* \qquad y_i^* = a + bx_i$$

Propriétés de la fonction des résidus

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i^*) = \bar{y} - \bar{y} = 0 \qquad \sum_{i=1}^{n} x_i e_i = 0 \qquad \sum_{i=1}^{n} y_i^* e_i = 0.$$

Régression simple : Analyse numérique

La somme des carrées

Régression SC : $\mathrm{SC}_{REGR} = \sum_{i=1}^{n} (y_i^* - \bar{y})^2$ *DL = p*

Erreur SC : $\mathrm{SC}_{RES} = \sum_{i=1}^{n} e_i^2$ *DL =n-p-1*

Somme des carrés totale : $\mathrm{SC}_{TOT} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ *DL =n-1*

$s = \sqrt{\mathrm{CME}}$ $\qquad s_e^2 = \frac{\mathrm{SC}_{RES}}{n} = \frac{1}{n} \sum_{i=1}^{n} e_i^2$

Décomposition de la variance

$s_{y^*}^2 = s_y^2 r^2$ $\qquad s_e^2 = s_y^2 (1 - r^2)$ $\qquad s_y^2 = s_{y^*}^2 + s_e^2.$

n nombre d'observations

p nombre de coefficients dans le modèle, constante non incluse

Qu'est-ce que le R carré ?

La valeur R^2 est le pourcentage de la variation de la variable de réponse expliqué par sa relation avec une ou plusieurs variables de prédiction. En général, plus la valeur R^2 est grande, plus le modèle est ajusté aux données. R^2 est toujours compris entre 0 et 100 %. Il est également appelé coefficient de détermination ou de détermination multiple (dans la régression linéaire multiple).

$$R^2 = 1 - \frac{\text{Erreur SC}}{\text{SC totale}}$$

Qu'est-ce que le R carré ajusté ?

Le R^2 ajusté est le pourcentage de la variation de la variable de réponse expliqué par la relation de cette variable avec une ou plusieurs variables de prédiction, ajusté en fonction du nombre de prédicteurs dans le modèle. Le R^2 ajusté permet de déterminer à quel point le modèle ajuste vos données lorsque vous souhaitez l'ajuster en fonction du nombre de prédicteurs inclus. La valeur du R^2 ajusté intègre le nombre de prédicteurs dans le modèle pour vous aider à choisir le modèle correct.

$$R^2_{\text{ajust}} = 1 - \frac{\text{Erreur CM}}{\text{Somme des carrés totale/DL total}}$$

Exemple

Step 1: Représentation graphique des data

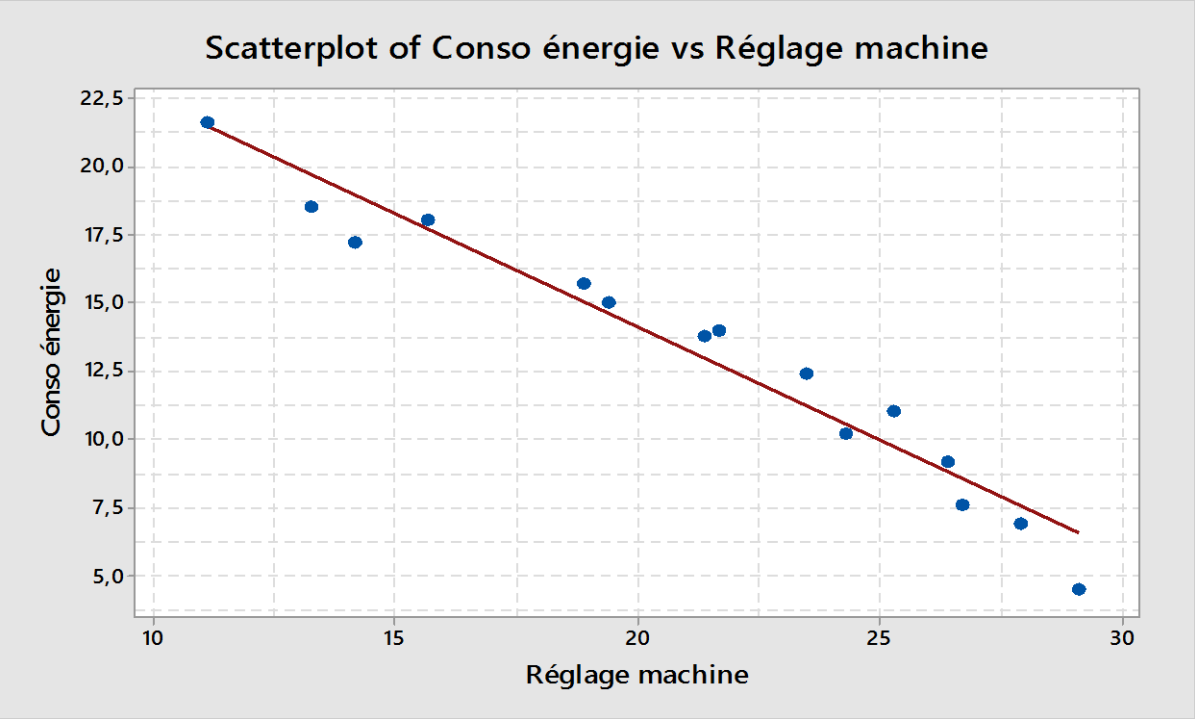
Step 2: Effectuer l'analyse de la corrélation

Step 3: Générer l'équation de prédiction

Step 4: Analyse du modèle

- Est ce qu'on a besoin d'un modèle a haut niveau?
- Vérifier la graphe du résidus

Step 1: Représentation graphique des data



Corrélation: Conso énergie; Réglage machine

Pearson corrélation of Conso énergie and Réglage machine = -0,974
P-Value = 0,000

Step 2: Effectuer l'analyse de la corrélation

P-value << 5% Une corrélation significative entre les ventes et les revenues

Step 3: Générer l'équation de prédiction

$$b = \frac{s_{xy}}{s_x^2} \quad et \quad a = \bar{y} - b\bar{x}$$

b = (371,71/15) / (5,46)² = - 0,83

a = 13,04 – 0,83 * 21,26 = 30,69

L'équation de régression

Conso énergie = 30,69 - 0,83 réglage machine

Regression Statistics	
Multiple R	-0,9744
R Square	94,95%
Adjusted R Square	94,56%
s	1,1238
Standard error	0,2902

Regression Equation:		
Intercept		30,69
slope	-	0,830

Regression Equation:

Conso énergie=30,69+-0,83*Réglage machine

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	308,60	308,60	244,37	0,0000
Residual	13	16,42	1,26		
Total	14	325,02			

Méthode matricielle

Cte	Réglage	Consommation
1	11,15	21,6
1	13,3	18,5
1	14,2	17,2
1	15,7	18
1	18,9	15,7
1	19,4	15
1	21,4	13,8
1	21,7	14
1	23,5	12,4
1	24,3	10,2
1	25,3	11
1	26,4	9,2
1	26,7	7,6
1	27,9	6,9
1	29,1	4,5
Moyenne	**21,26**	**13,04**
Ecart type	**5,46**	**4,65**

A'																
	Cte	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	Income Xi	11,15	13,3	14,2	15,7	18,9	19,4	21,4	21,7	23,5	24,3	25,3	26,4	26,7	27,9	29,1

A'A		
	15	318,95
	318,95	7229,66

$A'A^{-1}$		
	1,08	- 0,05
	- 0,05	0,00

A'y		
	195,6	195,6
	3787,4	3787,4

$(A'A)^{-1}*A'y$		
	30,693	*a*
	- 0,830	*b*

Conso énergie = 30,69 - 0,83 réglage machine

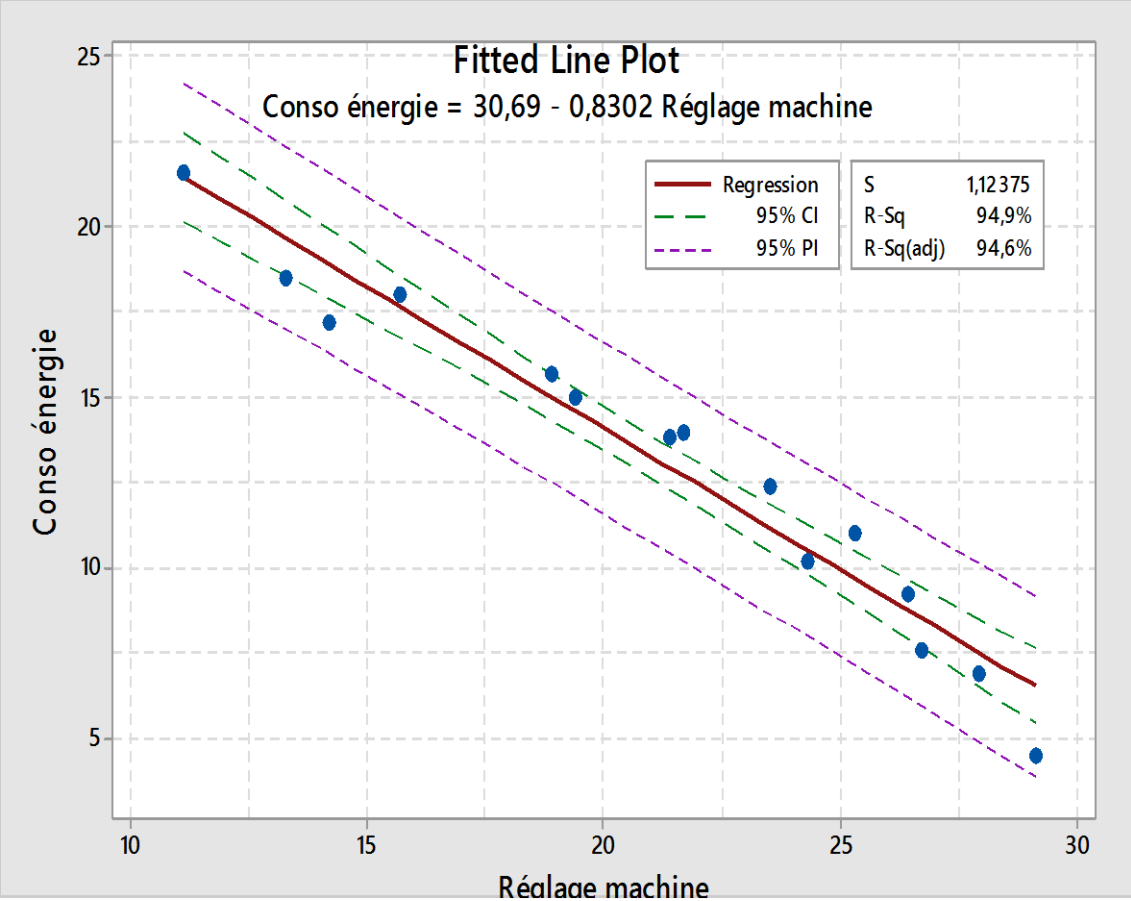
Pour la régression linéaire simple, l'erreur type du coefficient est la suivante

$$\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

$$t_i = \frac{\hat{\beta}_i}{\mathrm{ErT}(\hat{\beta}_i)}$$

Ti cal = -0,83 / (1,124/21,16) = 15,632 p = 0,000
Tcrit (0,025-13) = 2,16

Exemple



Équation de prédiction avec la ligne de forme. Est ce que R-Sq and R-Sq(adj) sont trés différentes?

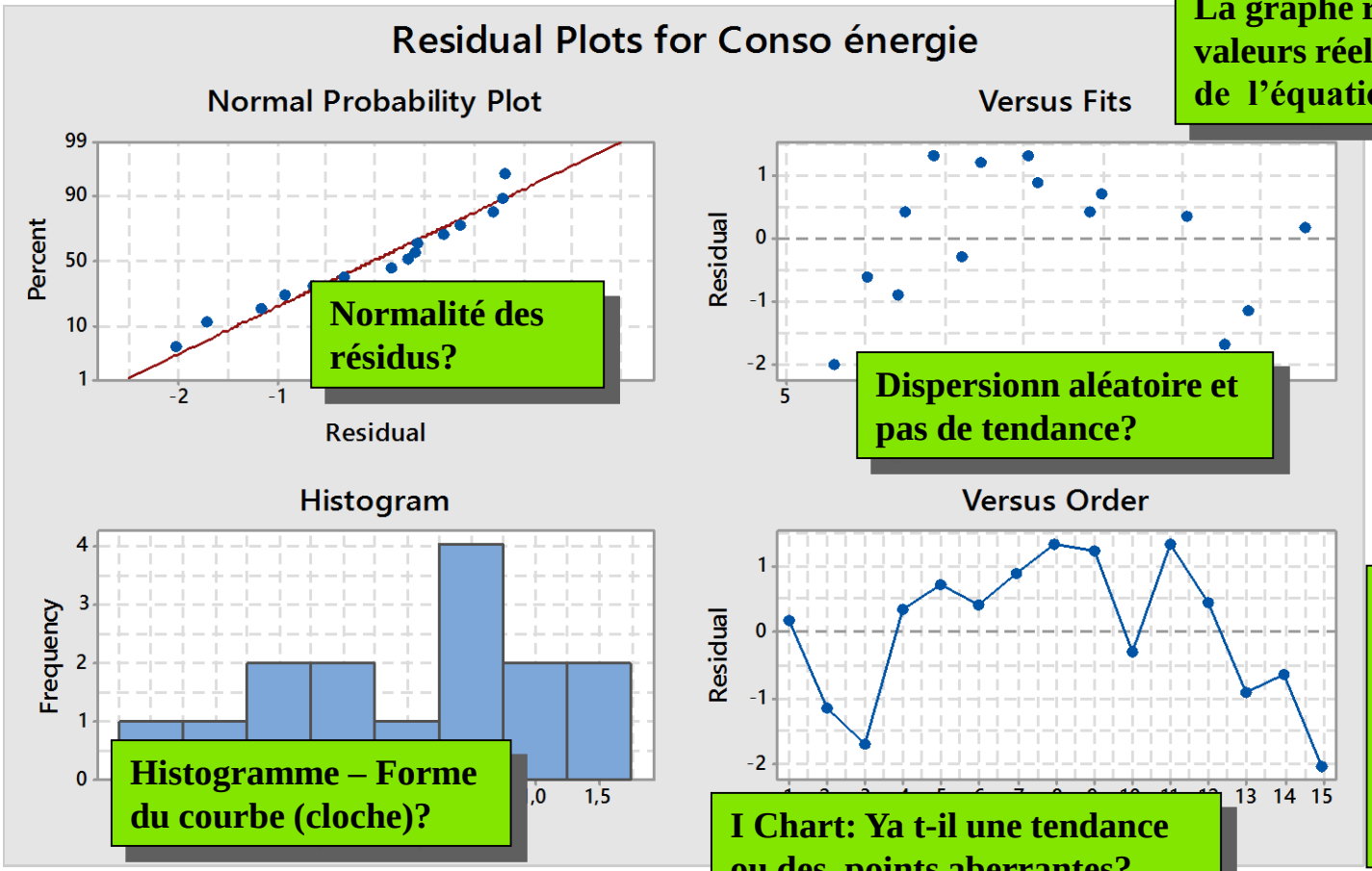
Toute valeur individuelle de la population, est entre les lignes bleu de prédiction bandes de confiance 95% .

La moyenne de tout échantillon de la population entre la ligne rouge bande de confiance 95.

Analyse des Résidus

Step 4: Analyse du modèle

- Vérifier la graphe des résidus.
- Courbe? Résidus?



La graphe représente la position des valeurs réelles par rapport aux valeurs de l'équation

Normalité des résidus?

Dispersionn aléatoire et pas de tendance?

Histogramme – Forme du courbe (cloche)?

I Chart: Ya t-il une tendance ou des points aberrantes?

La graphe présente comment le résidus se comporte le long de l'expérience.

La présentation doit être aléatoire

Résumé

- **L'analyse de la régression recherche une relation entre les variables sous la forme d'une équation de prédiction qui peut être ou non linéaire.**
- **Dans la régression, l'équation peut être soit la réponse souhaitée soit le moyen de prédiction désiré.**

Rationalisation des Tolérances

Rationalisation des Tolérances

- **Une méthode graphique simple pour établir les niveaux optimum et les tolérances appropriées des ENTRÉES.**

- **Dès que l'on a déterminé qu'une sortie continue dépend linéairement d'une entrée continue, la spécification des sorties est utilisée pour créer la spécification des entrées.**

- **Les nuages de points et les droites d'ajustement démontrent la relation entre les entrées et les sorties, mais pas nécessairement les causes et effets.**

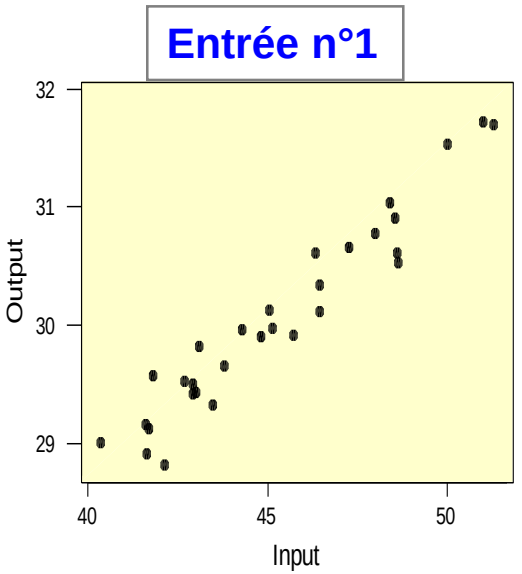
Rationalisation des Tolérances

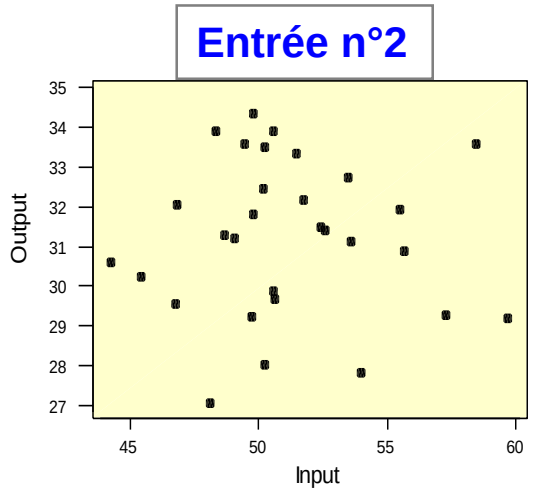
<u>Étape 1 :</u> Sélectionnez la variable de réponse et sa valeur cible. Ici, la spécification de la réponse est 30.5 +/- 1.0.

<u>Étape 2 :</u> Sélectionnez une variable d'entrée (entrée 1) avec une étendue optimale se situe entre 40 et 50. Une nouvelle variable d'entrée (entrée n°2) avec une étendue 45 à 60).

<u>Étape 3 :</u> Lancez 30 échantillons et mesurez le niveau de l'entrée et la sortie observée.

<u>Étape 4 :</u> Reportez les résultats dans un diagramme avec la variable d'entrée sur l'axe x et la sortie sur l'axe y. Si le diagramme a une pente avec un léger nuage de points vertical, il existe une relation. Passez à l'étape 5. S'il n'y a pas de pente, il n'existe aucune relation entre la variable d'entrée et la variable de réponse





Rationalisation des Tolérances

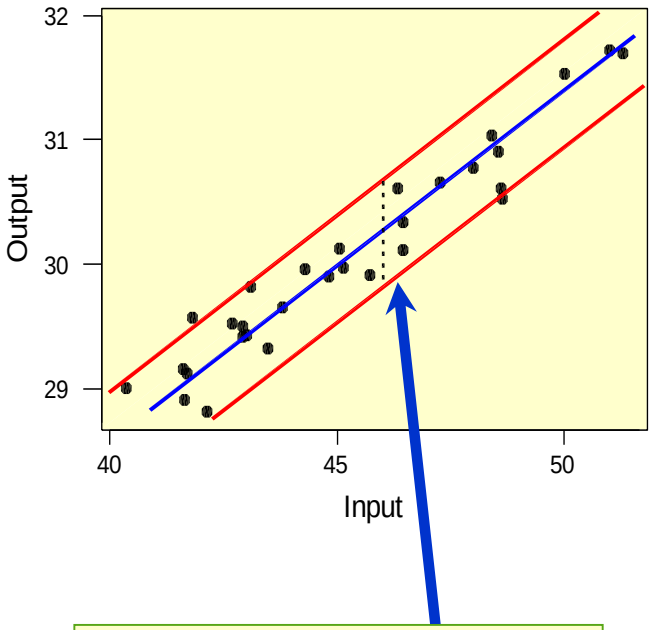
Étape 5 :

a) **Tracez la meilleure droite d'ajustement à travers les données.**

b) **Éliminez les points de données les plus éloignés de la meilleure droite d'ajustement.**

c) **Dessinez une droite parallèle passant par les prochains points les plus éloignés de la meilleure droite d'ajustement.**

d) **Tracez une seconde droite parallèle équidistante à la meilleure droite d'ajustement de l'autre côté. La distance verticale entre ces deux droites parallèles représente 95 % de l'effet total de tous les autres facteurs sur la sortie autres que la variable d'entrée étudiée ici.**



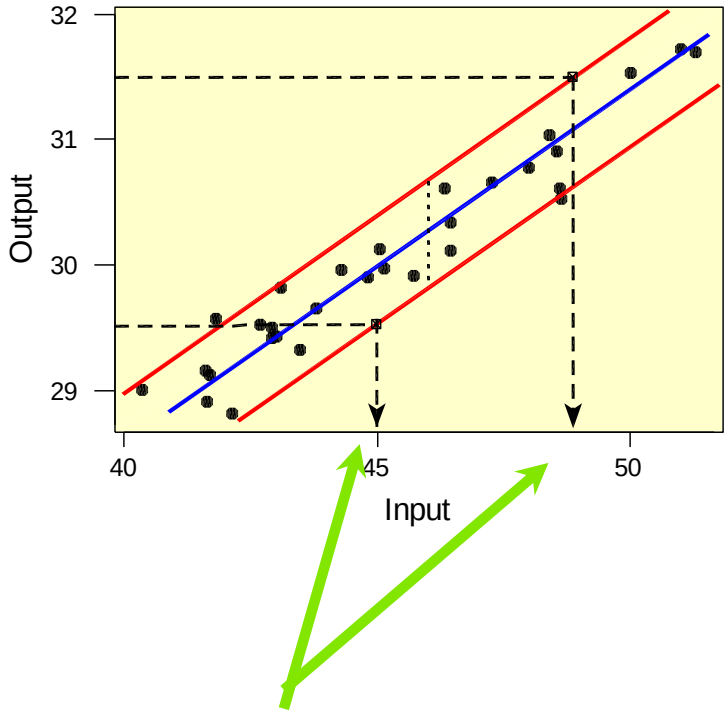
95 % de l'effet total des facteurs autres que cette variable d'entrée

Rationalisation des Tolérances

Étape 5 (suite) :

e) S'il existe des spécifications pour la variable de réponse, tracez les droites de ces valeurs sur l'axe y pour qu'elles croisent les droites de confiance supérieure et inférieure.

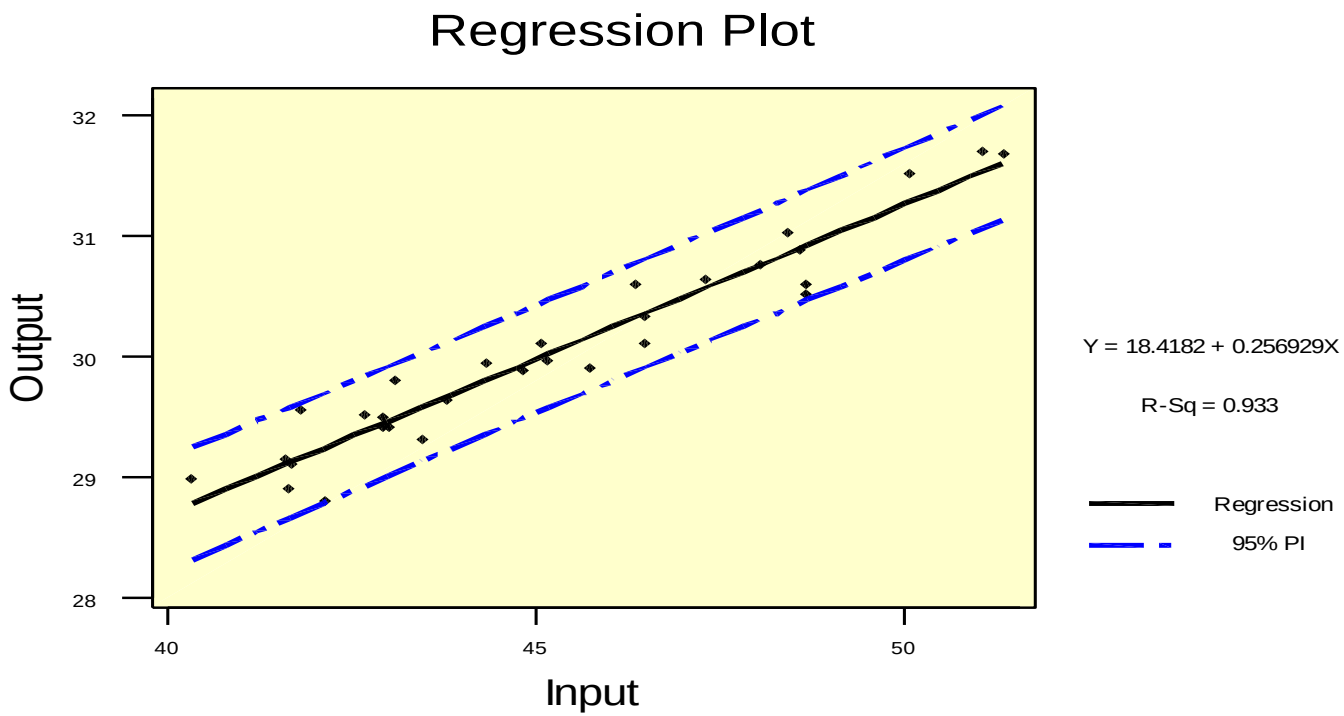
f) Faites descendre deux droites vers l'axe x à partir de ces points d'intersection. La distance entre les points où ces droites croisent l'axe x représente la tolérance maximum admissible de la variable d'entrée.



Tolérance de la variable d'entrée

Inf. = 45 et Sup. = 48.5

Rationalisation des Tolérances



Étape 6 : Comparez ces valeurs avec les niveaux de fonctionnement existants et implémentez les modifications nécessaires de la procédure standard d'exploitation. Documentez les modifications par la FMEA et le plan de contrôle.

Régressions linéaires multiples

Régressions linéaires multiples

L’analyse par régression linéaire multiple est une des solutions qui existe pour observer les liens entre une variable quantitative dépendante et *n* variables quantitatives indépendantes..

Trois paramètres doivent être vérifiés pour assurer la puissance du modèle obtenu:

- **L'échantillon doit être assez grand pour conférer suffisamment de puissance au test et pour fournir une précision suffisante pour l'estimation de l'importance de la relation entre X et Y.**
- **Il est important d'identifier les données aberrantes susceptibles d'influer les résultats de l'analyse.**
- **Le terme d'erreur doit suivre une loi Normale (analyse du résidu).**

Régressions linéaires multiples

Les équations se compliquent avec plusieurs variables, deux méthodes distinctes permettent de résoudre les équations.

- **La première repose sur la connaissance des coefficients de corrélation linéaire simple de toutes les paires de variables entre elles, de la moyenne arithmétique et des écarts-types de toutes les variables.**
- **La seconde repose sur des calculs matriciels.**

Régressions linéaires multiples

La formule du coefficient de corrélation

$$Ceof.cor = \frac{CovXY}{\sqrt{VarX} * \sqrt{VarY}}$$

Pour une régression multiple , l'analyse est plus compliquée qu'une régression simple mais suit la même logique

Linéaire	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

Les valeurs des coefficients β sont inconnues et doivent être estimées à partir des données. La méthode d'estimation est celle des moindres carrés, qui minimise la somme des valeurs résiduelles dans l'échantillon :

$$\min \sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i)\right)^2.$$

Une valeur résiduelle correspond à la différence entre la réponse observée Yi et la valeur ajustée $f(Xi)$ en fonction des coefficients estimés. La valeur minimisée de cette somme des carrés est la SCE (somme des carrés d'erreur) d'un modèle donné.

Multi-colinéarité dans la régression

Une multi-colinéarité prononcée s'avère problématique, car elle peut augmenter la variance des coefficients de régression et les rendre instables et difficiles à interpréter. Les conséquences de coefficients instables peuvent être les suivantes :

- **Les coefficients peuvent sembler non significatifs, même lorsqu'une relation significative existe entre le prédicteur et la réponse.**
- **Les coefficients de prédicteurs fortement corrélés varieront considérablement d'un échantillon à un autre.**
- **Lorsque des termes d'un modèle sont fortement corrélés, la suppression de l'un de ces termes aura une incidence considérable sur les coefficients estimés des autres. Les coefficients des termes fortement corrélés peuvent même présenter le mauvais signe.**

Multi-colinéarité dans la régression

Déterminer si les facteurs Xi sont fortement inter-corrélés
Dans une régression, la multi-colinéarité est un problème qui survient lorsque certaines variables de prévision du modèle sont corrélées avec d'autres.
Le FIV peut être obtenu en faisant régresser chaque prédicteur sur les prédicteurs restants et en notant la valeur R^2.

- **Chaque X a son Variance Inflation Factor (VIF)**

$$VIF_i = \frac{1}{1 - R_i^2}$$

$$R_i^2 = \text{R}^2 \text{ valuer obtenue en ajustant } \text{X}_\text{i} \text{ aux autre X's}$$

- **VIF est une règle statistique**
  - Si est supérieur à 7, la forme du modèle doit être modifiée

Valeurs aberrantes

Effets de levier (Hi)

Les effets de levier sont obtenus à partir de la matrice H, qui est une matrice de projection *n* x *n* :

$$H = X(X'X)^{-1}X'$$

L'effet de levier de la *i*e observation est le *i*e élément diagonal, h_i de H. Si la valeur de h_i est élevée, la *i*e observation contient des prédicteurs aberrants (X_{1i}, X_{2i}, ..., X_{pi}). En d'autres termes, les valeurs de prédicteurs sont éloignées du vecteur de moyenne $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_p)$, avec distance de Mahalanobis.

Les valeurs à effet de levier sont comprises entre 0 et 1. Dans le tableau des observations aberrantes, Les observations qui présentent des effets de levier supérieurs à 3 p/n.

Valeurs aberrantes =3*4/30 =0,4

Une régression linéaire multi-variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \text{-- } \text{-- } + \beta_p X_p$$

Y		ε		X				β			
Y_1	=	β_0	1	$X_{1.1}$	$X_{2.1}$	-	$X_{p.1}$	β1	β2	-	βp
Y_2			1	$X_{1.2}$	$X_{2.2}$	-	$X_{p.2}$	β1	β2	-	βp
Y_3			1	$X_{1.3}$	$X_{2.3}$	-	$X_{p.3}$	β1	β2	-	βp
-				-	-	-	-	-	-	-	-
-				-	-	-	-	-	-	-	-
Yn			1	$X_{1.n}$	$X_{2.n}$	-	$X_{p.n}$	β1	β2	+	βp

Exemple:

Une analyse de la distance du freinage d'une voiture par rapport aux facteurs indépendants:

- **La vitesse de la voiture 'speed'**
- **L'état des pneus 'Tirecond'**
- **Le temps de réaction 'Reactime'**
- **L'état de la route 'Strtcond'**

Soit r12, r13… rpp les coefficients de corrélations linéaires des paires de variables et s1, s2, …, sp les écarts-types

Y	Xi			
Brkleng	Speed	Tirecond	Reactime	Strtcond
25,74	44,90	2,01	0,87	7,82
24,75	49,49	2,60	0,23	6,19
28,82	48,19	2,82	0,43	3,71
20,46	42,14	2,21	0,45	8,10
23,03	45,61	2,16	0,18	6,96
37,63	53,61	1,64	1,32	4,99
21,68	38,28	3,63	1,27	4,37
28,53	45,41	2,62	1,00	5,82
28,85	52,67	3,68	1,68	7,36
32,38	53,32	4,02	1,11	4,62
41,93	63,25	1,96	0,77	3,93
38,62	53,14	2,17	1,71	5,07
28,47	48,88	2,34	0,76	5,95
35,49	51,80	3,06	1,12	4,07
31,95	55,73	2,97	0,04	2,39
27,39	47,86	2,66	1,02	7,24
26,89	51,45	3,26	0,53	6,80
31,28	51,32	2,86	1,36	7,89
30,76	53,93	3,15	1,37	7,44
24,22	45,76	4,05	1,17	6,25
42,06	58,09	2,64	1,49	2,00
26,56	49,09	3,57	0,25	5,09
26,04	47,32	3,61	0,65	3,18
26,22	50,26	2,62	0,20	4,30
36,18	51,13	1,78	1,12	4,64
27,87	47,77	3,73	1,27	4,63
31,22	51,67	2,83	1,63	7,48
28,53	47,67	2,71	0,97	6,07
36,91	52,72	2,37	1,38	5,01
34,33	53,13	2,24	0,70	5,89
30,16	50,19	2,80	0,93	5,51

Calcul Matricielle

1- On ajoute un vecteur unitaire V

Y		Xi			
Brkleng	V	Speed	Tirecond	Reactime	Strtcond
25,74	1	44,90	2,01	0,87	7,82
24,75	1	49,49	2,60	0,23	6,19
28,82	1	48,19	2,82	0,43	3,71
20,46	1	42,14	2,21	0,45	8,10
23,03	1	45,61	2,16	0,18	6,96
37,63	1	53,61	1,64	1,32	4,99
21,68	1	38,28	3,63	1,27	4,37
28,53	1	45,41	2,62	1,00	5,82
28,85	1	52,67	3,68	1,68	7,36
32,38	1	53,32	4,02	1,11	4,62
41,93	1	63,25	1,96	0,77	3,93
38,62	1	53,14	2,17	1,71	5,07
28,47	1	48,88	2,34	0,76	5,95
35,49	1	51,80	3,06	1,12	4,07
31,95	1	55,73	2,97	0,04	2,39
27,39	1	47,86	2,66	1,02	7,24
26,89	1	51,45	3,26	0,53	6,80
31,28	1	51,32	2,86	1,36	7,89
30,76	1	53,93	3,15	1,37	7,44
24,22	1	45,76	4,05	1,17	6,25
42,06	1	58,09	2,64	1,49	2,00
26,56	1	49,09	3,57	0,25	5,09
26,04	1	47,32	3,61	0,65	3,18
26,22	1	50,26	2,62	0,20	4,30
36,18	1	51,13	1,78	1,12	4,64
27,87	1	47,77	3,73	1,27	4,63
31,22	1	51,67	2,83	1,63	7,48
28,53	1	47,67	2,71	0,97	6,07
36,91	1	52,72	2,37	1,38	5,01
34,33	1	53,13	2,24	0,70	5,89
30,16	1	50,19	2,80	0,93	5,51

2- Faire la matrice transposé X' de X

X'																														
	V	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	Speed	44,90	49,49	48,19	42,14	45,61	53,61	38,28	45,41	52,67	53,32	63,25	53,14	48,88	51,80	55,73	47,86	51,45	51,32	53,93	45,76	58,09	49,09	47,32	50,26	51,13	47,77	51,67	47,67	52,72	53,13	50,19
	Tirecond	2,01	2,60	2,82	2,21	2,16	1,64	3,63	2,62	3,68	4,02	1,96	2,17	2,34	3,06	2,97	2,66	3,26	2,86	3,15	4,05	2,64	3,57	3,61	2,62	1,78	3,73	2,83	2,71	2,37	2,24	2,80
	Reactime	0,87	0,23	0,43	0,45	0,18	1,32	1,27	1,00	1,68	1,11	0,77	1,71	0,76	1,12	0,04	1,02	0,53	1,36	1,37	1,17	1,49	0,25	0,65	0,20	1,12	1,27	1,63	0,97	1,38	0,70	0,93
	Strtcond	7,82	6,19	3,71	8,10	6,96	4,99	4,37	5,82	7,36	4,62	3,93	5,07	5,95	4,07	2,39	7,24	6,80	7,89	7,44	6,25	2,00	5,09	3,18	4,30	4,64	4,63	7,48	6,07	5,01	5,89	5,51

3-Faire le produit des matrice X'X

X'X					
	31,00	1 555,78	86,77	28,98	170,77
	1 555,78	78 747,66	4 335,34	1 465,97	8 489,90
	86,77	4 335,34	255,67	82,04	474,91
	28,98	1 465,97	82,04	33,89	162,79
	170,77	8 489,90	474,91	162,79	1 019,23

4-Chercher la matrice inverse $(X'X)^{-1}$

(X'X)-1					
	8,27	- 0,12	- 0,46	0,22	- 0,22
	- 0,12	0,00	0,00	- 0,00	0,00
	- 0,46	0,00	0,09	- 0,02	0,01
	0,22	- 0,00	- 0,02	0,16	- 0,01
	- 0,22	0,00	0,01	- 0,01	0,02

5-Faire le produit des matrices X'Y

X'Y	
	934,95
	47 587,60
	2 576,26
	909,30
	5 028,63

6- Faire le produit des matrices $(X'X)^{-1}X'Y$

$(X'X)^{-1}*X'y$		
	4,60	Cte
	0,696	b1
	- 2,749	b2
	4,907	b3
	- 1,140	b4

Analyse du modèle

Tableau Anova

Source	DL	SS	MS	F	P
Regression	4	886,49	221,62	134,58	0,0000
Speed	1	245,59	245,59	149,13	0,0000
Tirecnd	1	86,37	86,37	52,45	0,0000
Reactime	1	146,60	146,60	89,02	0,0000
Strtcond	1	81,33	81,33	49,39	0,0
Erreur	25	41,17	1,65		
Total	29	927,7			

SS régression $\sum (\hat{y}_i - \bar{y})^2$

est la part de la variation expliquée par le modèle

SS erreur $\sum (y_i - \hat{y}_i)^2$

est la part de la variation qui n'est pas expliquée par les données

SS total $\sum (y_i - \bar{y})^2$

$\hat{Y}_i$: Valeur modélisée

Calcul des sommes des carrées des facteurs:
La SS est calculé pas à pas en éliminant chaque fois un facteur et en recalculant la nouvelle SS erreur puis en déduisant la valeur de l'erreur pure

Les erreurs types des coefficients pour la régression multiple sont les racines carrées des éléments de diagonale de la matrice :

ErT = $(X'X)^{-1}s^2$ $\quad s = \sqrt{CME}$

(X'X)-1	8,27	- 0,12	- 0,46	0,22	- 0,22
	- 0,12	0,00	0,00	- 0,00	0,00
	- 0,46	0,00	0,09	- 0,02	0,01
	0,22	- 0,00	- 0,02	0,16	- 0,01
	- 0,22	0,00	0,01	- 0,01	0,02

S^2	Ert
1,65	3,691
1,65	0,057
1,65	0,380
1,65	0,520
1,65	0,162

Term	Coef	SECoef	T-Value	Pvalue
Cste	4,60	3,69	1,247	0,2245
Speed	0,70	0,06	12,212	0,0000
Tirecond	- 2,75	0,38	- 7,242	0,0000
Reactime	4,91	0,52	9,435	0,0000
Strtcond	- 1,14	0,16	- 7,027	0,0000

$$t_i = \frac{\hat{\beta}_i}{ErT(\hat{\beta}_i)}$$

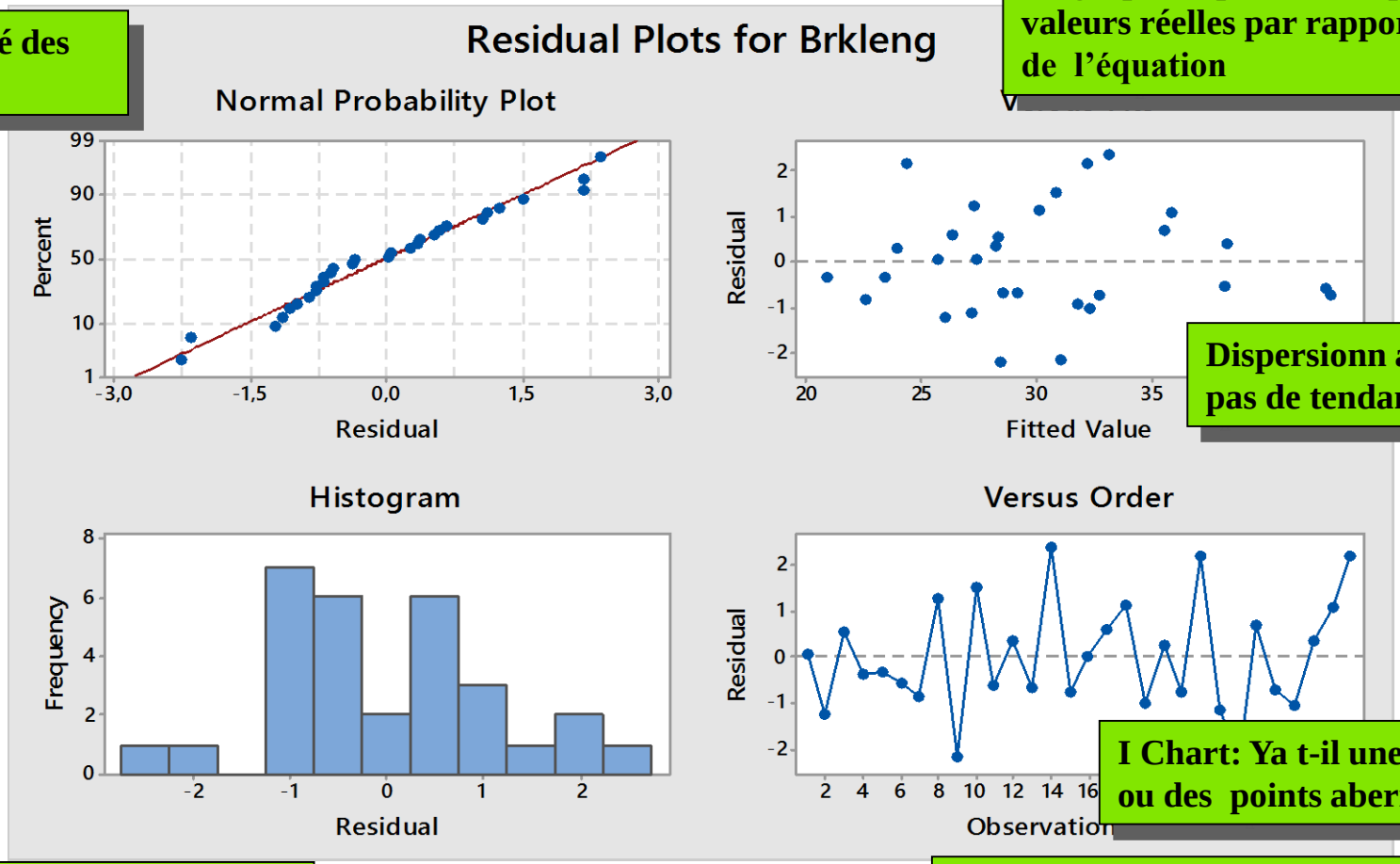
DL = n-1-p

P = nbre des facteurs

Analyse des résidus

Normalité des résidus?

La graphe représente la position des valeurs réelles par rapport aux valeurs de l'équation



Dispersionn aléatoire et pas de tendance?

I Chart: Ya t-il une tendance ou des points aberrantes?

Histogramme – Forme du courbe (cloche)?

La graphe présente comment le résidus se comporte le long de l'expérience.

La présentation doit être aléatoire

Questions?