UN GROUPE D'INDIVIDUS DÉCRIT PAR DEUX VARIABLES QUANTITATIVES

PRÉDIRE UNE VARIABLE À PARTIR DE L'AUTRE

Mots-clés : Variables numériques ; Prédiction ; Régression ; Erreur ; Résidu ; Valeurs prédites ; Valeurs résiduelles.

Ce document a été établi en indiquant comment obtenir les différents résultats avec le logiciel SES-Pegase (version 7). Cependant, il peut être utilisé comme guide méthodologique et d'interprétation, quel que soit le logiciel utilisé.

TYPE DES DONNÉES ANALYSÉES

Nous présenterons l'analyse d'un dossier particulier, le dossier CLIMAT. Mais cette analyse s'applique à toutes données de la forme suivante (Tableau 1) :

Tableau 1 : Structure du tableau des données individuelles

INDIV	X1	X2
i1	10	300
i2	12	50
i3	20	31
i4	7	63
i5	8	71
i6	20	18
i7	1	100
i8	18	103
(…)	(…)	(…)

On a recueilli des informations sur un ensemble d'individus (i1, i2…). Ces individus peuvent être, des personnes, des pays, des animaux, des voitures…
Parmi les données recueillies, on a deux variables quantitatives observées sur les individus : leur note, leur temps de réponse, leur âge, leur taille, leur poids…

Les deux variables ne sont pas nécessairement dans la même unité.
L'une des deux variables a le statut de variable prédictrice (VP), l'autre de variable à prédire ou variable dépendante (VD).

QUESTION

Est-il possible de trouver une équation simple qui permette de prédire, au moins en partie, les valeurs d'une variable (VD) connaissant les valeurs de l'autre variable (VP) ? Plus précisément, on se demande :

1. Quelle est la forme de la liaison entre les deux variables ? Est-elle de type linéaire, ou autre ?
2. Si la liaison est de type linéaire, quelle équation permet de faire des prédictions ?
3. Quel est le sens de la liaison linéaire, positif ou négatif ?
4. Quelle est sa force de la liaison linéaire, faible ou forte ?
5. Quelle est la qualité de la prédiction linéaire ?

UN EXEMPLE : LE DOSSIER CLIMAT

Les données[1]

On dispose dans le dossier CLIMAT de deux types d'informations concernant 12 pays (Allemagne, Autriche, Belgique, Finlande, Norvège, Royaume-Uni, Suède, Espagne, Italie, Portugal, Hongrie, Pologne).

1. Les données de l'OMS (Organisation Mondiale de la Santé) qui indiquent, pour l'année 1999, le taux de suicides, global, par sexe et par tranche d'âge :
- TSUIC : Le taux de suicide total (pour 100.000 habitants)
- TSHOM : Le taux de suicide chez les Hommes (pour 100.000)
- TSFEM : Le taux de suicide chez les Femmes (pour 100.000)
- TS1534 : Le taux de suicide chez les 15-34 ans (pour 100.000)
- TS3554 : Le taux de suicide chez les 35-54 ans (pour 100.000)
- TS5574 : Le taux de suicide chez les 55-74 ans (pour 100.000)

2. Les données de l'INED (Institut National des Études Démographiques) qui indiquent :
- TEMPT : La température moyenne (en degrés Celsius) sur l'année
- PLUIE : La quantité de précipitations (pluie + neige) en un an (en litres/m²).

On s'intéresse ici uniquement à la liaison entre deux variables :
- la température moyenne (TEMPT),
- le taux de suicide total (TSUIC).

Tableau 2 : Données CLIMAT

PAYS	TSUIC	TEMPT	PLUIE	TSHOM	TSFEM	TS1534	TS3554	TS5574
ALLE	11.38	8.5	583	8.55	2.83	2.59	4.56	4.22
AUTR	15.38	9.1	684	11.49	3.88	3.74	6.22	5.42
BELG	17.30	9.7	833	12.55	4.75	4.82	7.50	4.98
FINL	21.32	4.5	605	16.07	5.25	6.34	10.16	4.83
NORV	12.22	4.2	885	9.09	3.13	4.97	4.41	2.83
ROYA	6.81	10.2	599	5.37	1.44	2.50	2.91	1.40
SUED	11.78	5.9	569	8.32	3.46	2.91	5.17	3.69
ESPA	6.49	14.4	492	4.90	1.60	1.78	2.24	2.48
ITAL	5.88	15.7	828	4.45	1.43	1.56	1.93	2.39
PORT	3.83	16.0	682	3.08	0.75	0.83	1.21	1.79
HONG	25.18	10.9	596	19.85	5.33	4.51	12.67	8.00
POLO	14.24	7.6	511	11.97	2.27	3.95	7.23	3.06

Type et statut des variables

1. Ces deux variables ne sont pas sur la même échelle une proportion (sur mille habitants) et une température (en degrés Celsius). Mais elles sont, pour l'essentiel, de même type : ce sont deux variables quantitatives. Notons que la variable TSUIC est, plus précisément, une variable de rapport (cas particulier de variable quantitative). On mentionnera dans l'analyse quelques spécificités liées à ce type de variable.

2. Les deux variables ont des statuts différents :
- le taux de suicide (TSUIC) a, compte tenu de la question posée, le statut de variable à prédire ou variable dépendante (VD),
- la température moyenne (TEMPT) a le statut de variable prédictrice ou variable indépendante (VI).

1 Sources : http://www5.who.int/mental_health et http://www.ined.fr

Questions

Est-il possible de trouver une équation simple qui permette de **prédire**, au moins en partie, le taux de suicide d'un pays (TSUIC) connaissant la température moyenne (TEMPT) de ce pays ? Plus précisément, on se demande :

1. Quelle est la forme de cette liaison ? Est-elle de type linéaire, ou autre ?
2. Si la liaison est de type linéaire, quelle équation linéaire permet de faire des prédictions ?
3. Quel est le sens de la liaison linéaire, positif ou négatif ?
4. Quelle est sa force de la liaison, faible ou forte ?
5. Quelle est la qualité de la prédiction ?

Ouvrir le fichier

```
SES-Pégase
Lancer SESAnalyse
Menu Fichier
- Ouvrir un dossier (*.SES)
Sélectionner le fichier CLIMAT.SES Ouvrir
```

ANALYSER LES VARIABLES UNE À UNE

Avant de répondre à la question posée, on commencera par analyser chacune de ces deux variables séparément : distribution, tendance centrale, dispersion.
On se reportera à la présentation du dossier NOTEBAC pour une analyse détaillée d'une variable quantitative.

Analyser la VD (TSUIC)

```
Menu Nouvelle analyse
Sélectionner la variable TSUIC comme "Variable(s) à analyser" OK
```

Forme de la distribution ?

```
Menu Statistiques
- Distribution
– Histogramme
```

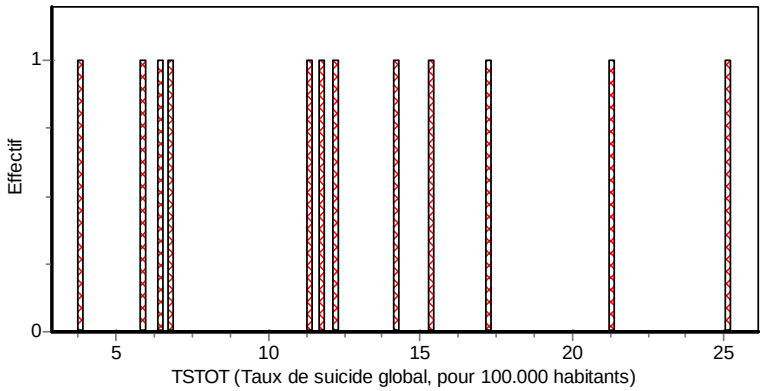

Figure 1 : Représentation graphique (histogramme) de la distribution des taux de suicide (TSUIC).

La distribution suggère, là aussi, l'existence de deux groupes de pays selon leur taux de suicide, un groupe minoritaire (4 pays) avec un taux de suicide relativement faible (autour de 5 pour 100.000 habitants) et un groupe majoritaire (8 pays) où ce taux est plus élevé (de 11.4 à 25.2 pour 100.000 habitants).
L'examen du tableau de données montre que le taux de suicide le plus faible est celui du Portugal. Le plus élevé est celui de la Hongrie.

Tendance centrale ?

```
Menu Statistiques
- Tendance centrale
- Tous indices de tendance centrale
```

Tableau 3 : Moyenne, médiane et mode des taux de suicide (pour 100.000 habitants)

	TSUIC
Moy	12.7
Med	11.8
Mod	Erreur

La moyenne des taux de suicide est d'environ 13 pour 100.000 habitants (moy=12.7). Le taux de suicide médian est de d'environ 12 (med = 11.8). Comme pour l'autre variable il n'existe pas de valeur modale, d'où le message "Erreur" pour le Mode).

Dispersion ?

```
Menu Statistiques
- Dispersion
– Quartiles
```

Tableau 4 : Min, max et répartition en quartiles des taux de suicide (pour 100.000 habitants) par pays

	Min	Q1	Med	Q3	Max
TSUIC	3.83	6.65	12	16.3	25.2

Pour ces 12 pays le taux de suicide varie de 4 à 25 pour 100.000 habitants (min = 3.83, max = 25.2).
La moitié des pays ont un taux de suicide compris entre 7 et 16 pour 100.000 habitants (Q1 = 6.65, Q3 = 16.3).

Cette variable étant une variable de rapport, cela permet de faire le rapport entre les maximum et minimum. Ce rapport (25.2 / 3.83 = 6.6) indique que le taux de suicide le plus élevé, celui de la Hongrie, est presque sept fois plus élevé que le plus faible, celui du Portugal.

Analyser la VI (TEMPT)

```
Menu Nouvelle analyse
Sélectionner la variable TEMPT comme "Variable(s) à analyser"
```

Forme de la distribution ?

```
Menu Statistiques
– Distribution
- Histogramme
```

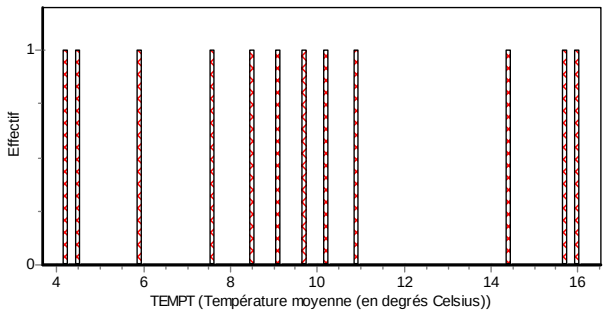

Figure 2 : Représentation graphique (histogramme) de la distribution des températures moyennes (TEMPT).

La distribution suggère l'existence de deux groupes de pays selon leur température moyenne, un groupe principal (9 pays) où les températures moyennes sont plus faibles (de 4° à 11°), et un groupe minoritaire (3 pays) où les températures sont plus élevées (de 14° à 16°). L'examen du tableau de données montre que ces trois pays sont l'Espagne, l'Italie et le Portugal.

Tendance centrale ?

```
Menu Statistiques
– Tendance centrale
- Tous indices de tendance centrale
```

Tableau 5 : Moyenne, médiane et mode des températures (en °C)

	TEMPT
Moy	9.7
Med	9.4
Mod	Erreur

On se demandera si les indices de tendance centrale sont proches ou, au contraire, divergent beaucoup. Dans ce dernier cas il faudra s'interroger sur les raisons de cette divergence : dissymétrie de la distribution ? Valeur(s) atypique(s) ?

La moyenne des températures moyennes de ces 12 pays européens est d'environ 10° C. La température médiane est d'environ 9°C. Le message "Erreur" s'affiche car il n'y a pas de valeur modale (cf. Figure 2).
La valeur plus élevée de la moyenne (9.7) par rapport à la médiane (9.4) est probablement due aux trois pays, relativement atypiques, avec des températures moyennes élevées (>14°C)..

Dispersion ?

```
Menu Statistiques
– Dispersion
– Quartiles
```

Tableau 6 : Répartition par quartiles des températures selon les pays (en °C)

	Min	Q1	Med	Q3	Max
TEMPT	4.2	6.75	9.4	12.7	16

On a vu précédemment que, pour ces 12 pays européens, la moyenne des températures est de presque 10°C. Toutefois les températures varient selon les pays. Elles sont comprises entre 4°C (min = 4.2) et 16°C (max = 16°) et 50% des pays ont une température moyenne comprise entre 7° et 13° (Q1 = 6.75, Q3= 12.7).

FORME DE LA LIAISON ?

On s'intéresse maintenant aux liens pouvant exister entre ces deux variables. On commence donc par sélectionner ces deux variables.

```
Menu Nouvelle analyse
Sélectionner TSUIC en tant que "Variable(s) à analyser - VD"
et TEMPT en tant que "Prédictrice(s) - VI".
```

Une des méthodes permettant d'évaluer la possibilité de prédire une variable à partir de l'autre est la régression linéaire. Afin de voir si cette méthode est applicable sur les données analysées, on va se demander si la liaison est bien, au moins en partie, de type linéaire. Pour cela on analyse le graphique bivarié pondéré.

```
Menu Statistiques
– Liaison VI*VD
– Graphique de corrélation
La taille des points peut être modifiée (cf. icônes au-dessus du graphique)
Voir aussi les autres options à droite du graphique
```

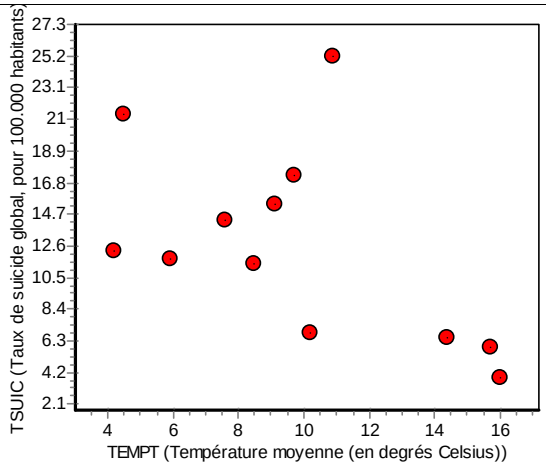

Figure 3 : Représentation graphique (nuage pondéré) de la liaison entre les deux variables

Le caractère linéaire de la liaison peut se discuter ici compte tenu de la forme du nuage de points. Cette impression est due en partie à l'existence d'un point (en haut à droite) qui semble atypique. Nous admettrons cependant que ce nuage peut être ajusté par une droite.

```
SES-Pegase
Approcher le curseur du point atypique : il s'agit de la Hongrie (HONG).
```

Ajustement du nuage par une droite ?

On représente graphiquement la droite qui ajuste au mieux ce nuage de points, selon un critère qui reste à préciser (cf. Figure 4). [2]

```
Menu Statistiques
- Régression Linéaire (RL)
- RL - Droite de régression
```

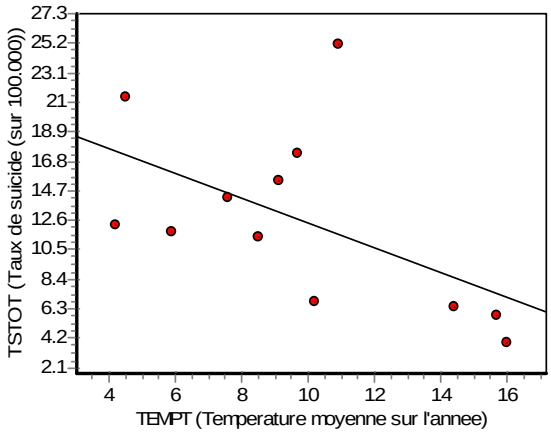

Figure 4 : Représentation graphique (nuage pondéré) de la droite de régression

Le fait que le logiciel nous présente une droite qui ajuste le nuage de points, n'est pas la preuve que cet ajustement est pertinent. En effet, à de rares exceptions près, il est toujours possible de trouver une telle droite. Des indices nous aideront à discuter de la pertinence de cet ajustement linéaire.

[2] Il existe plusieurs méthodes de régression pour calculer une équation linéaire. La méthode utilisée ici, la plus classique est désignée par méthode des Moindres Carrés Ordinaires (MCO) soit, en anglais, Ordinary Least Squares (OLS).

QUELLE ÉQUATION DE RÉGRESSION ?

```
Menu Statistiques
- Régression Linéaire (RL)
- RL - Coefficients de régression b0 et b
```

Tableau 7 : Coefficients de régression

	b0	TEMPT
TSUIC	21.2	-0.88

Au vu de ce tableau, l'équation de la droite de régression visant à prédire le taux de suicide (TSUIC) selon la température moyenne annuelle (TEMPT) est donc la suivante :

$TSUIC_{pred}$ = 21.2 - 0.88 x TEMPT

Il faut insister sur le fait que, quand bien même il existerait une liaison entre ces deux variables, et quand bien même on pourrait prédire précisément les valeurs d'une des variables connaissant les valeurs à l'autre variable, cela ne permet pas de conclure à l'existence d'un effet de la température sur le taux de suicide.
Ainsi, on constate un lien entre la consommation de glaces et le nombre de noyades sur les plages et il est possible de prédire l'une de ces variables à partir de l'autre. Mais cela ne permet pas de conclure à l'existence d'un effet de l'une sur l'autre. En l'occurrence cette liaison, et la possibilité d'une prédiction, sont dues à une troisième variable : le nombre de personnes présentes sur les plages, nombre lui-même dépendant de l'ensoleillement, de la température de l'eau…

Prédictions et résidus ?

Cette équation permet de prédire le taux de suicide dans un pays compte tenu de sa température moyenne. Ainsi pour l'Allemagne, où la température moyenne est égale à 8.5, le taux de suicide prédit par cette équation est :
$TSUIC_{pred}$ = 21.2 - 0.88 x 8.5 = 13.73 [3]

Il existe un écart entre le taux de suicide observé en l'Allemagne (11.38) et ce taux prédit (13.73). L'écart (observé - prédit) est de -2.35. On dénomme résidu ou erreur cet écart.

Faire des prédictions extérieures à l'échantillon ?

Le coefficient b0 ou ordonnée à l'origine.

Cette équation permet de faire une prédiction extérieure à l'échantillon : le coefficient b0 (+21.2) nous donne une prédiction immédiate (mais qui fait rarement sens). Il nous prédit que, dans un pays où la température moyenne annuelle (TEMPT) serait de 0°C, le taux de suicide (TSUIC) serait d'environ 21 (21.2) pour 100.000 habitants.

On utilisera l'équation de régression uniquement pour faire des prédictions à partir de valeurs proches de celles observées dans l'échantillon. Ainsi 0° n'est pas dans l'intervalle des températures moyennes observées dans l'échantillon. Il est donc hasardeux de faire une prédiction pour un pays où la température moyenne est proche de 0°C.

Autres prédictions

Supposons que l'on connaisse la température moyenne d'un pays, par exemple 12°C, mais que l'on ne connaisse pas son taux de suicide. Quel taux de suicide peut-on prévoir par la relation linéaire ?
L'équation prédit un taux de suicide de 10.6 suicides pour 100.000 habitants (21.2 - 0.88 × 12 = 10.6).

[3] *Pour retrouver cette valeur il faut utiliser les valeurs précises (21.1964…et -0.8787…) des coefficients et non pas les valeurs arrondies.*

SENS DE LA LIAISON LINÉAIRE ?

Dans l'échantillon ?

Le sens de la liaison est donné par le signe du coefficient de régression b (cf. Tableau 7) ou par la pente de la droite de régression (cf. Figure 4). On constate que le coefficient de régression (-0.88) est négatif et que la pente de la droite est également négative (c'est toujours le cas en régression simple).
Il apparaît donc que, globalement, la liaison linéaire est négative entre TEMPT et TSUIC : en moyenne, plus la température moyenne du pays augmente, plus le taux de suicide tend à diminuer.

Il est toujours possible, sauf rares exceptions, de trouver une équation linéaire permettant de prédire, pour un échantillon, les valeurs de la variable à prédire, connaissant les valeurs des prédictrices. La question qui se pose alors est de savoir s'il existe un lien dans la population d'où provient cet échantillon. Le test *F* vise à répondre à cette question.

Dans la population ?

La population est, de manière générale, l'ensemble d'où proviennent les individus de l'échantillon. Mais les notions d'individu et de population sont à prendre au sens statistique. Ici les individus sont des pays. Dès lors la population est l'ensemble des pays européens.

Conditions préalables à l'inférence?

1. La mise en œuvre des procédures inférentielles suppose que l'échantillon des 12 pays ait été tiré au hasard parmi l'ensemble des pays européens.
2. Une autre condition, pour la mise en œuvre d'un test t de Student, est la normalité de la distribution des erreurs de prédiction dans la population. Cette condition étant impossible à vérifier, du fait qu'on ne connait pas la population, on examinera ce qu'il en est dans l'échantillon.

```
Menu Statistiques
- Régression linéaire (RL)
- Histogramme des résidus
```

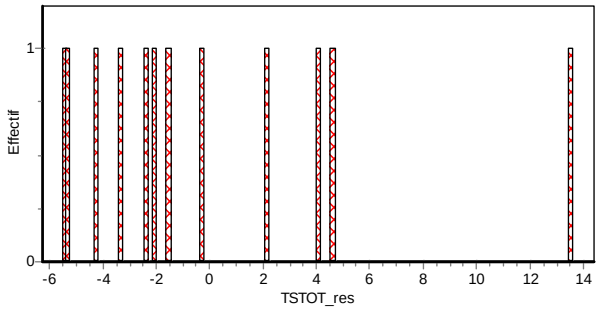

Figure 5 : Histogramme des résidus (TSUIC_res) de la régression

L'examen de la distribution des résidus (cf. Figure 5) montre qu'il existe une valeur très atypique. Ce constat peut amener à mettre en doute l'hypothèse de normalité de la distribution des résidus dans la population. Dans cas, la validité du test t (cf. **Erreur ! Source du renvoi introuvable.**) serait à remettre en cause.

Test t de Student

```
Menu Statistique
- Régression linéaire (RL)
- RL - Test t (existence d'un effet de chaque VI)
```

Tableau 8 : Coefficient b et test t.

Variables	Beta	Erreur-ty	b	Erreur-ty	t	ddl	p
TEMPT	-0.54	0.27	-0.88	0.43	-2.05	10	0.0672

L'interprétation dépend de la valeur *p*, que l'on compare un une valeur repère conventionnelle (.05) :

Si $p < .05$ (ce n'est pas le cas ici), on déclare que "le test est significatif". Ce résultat permet de conclure que, dans la population, le coefficient *b* (et donc la liaison linéaire) est de même sens que dans l'échantillon.
Si $p > .05$ (c'est le cas ici) on dit que "le test est non significatif". Dans ce cas, on ne peut pas conclure sur le sens de la liaison linéaire dans la population.

Ici, on a $p = .0672$ ($p > .05$). On ne sait donc pas si, dans l'ensemble des pays, il existe ou non un lien entre la température moyenne (TMPT) et le taux de suicide (TSUIC) ou, dit autrement, si la température moyenne d'un pays permet de prédire son taux de suicide.

Compte tenu du test non significatif, la seule chose qu'on peut conclure est… qu'on ne sait pas !
Attention : le résultat de ce test, non significatif, ne permet PAS d'affirmer qu'il n'y a pas de liaison entre la température d'un pays et son taux de suicide.
Il faut clairement dissocier la conclusion du test (significatif ou non significatif) de la question de la force de l'effet. Un test significatif est compatible avec un effet faible. Un test non significatif ne permet pas de conclure à un effet faible, ni a fortiori à un effet nul.

FORCE DE LA LIAISON LINÉAIRE

Pour évaluer la force de la liaison linéaire, il existe deux types d'indices :
- des indices bruts (dépendants de l'échelle de mesure), par exemple le coefficient de régression,
- des indices calibrés (indépendants de l'échelle de mesure) tels que le coefficient de corrélation linéaire de Bravais Pearson (*r*).

Dans l'échantillon ?

Coefficient de régression ?

Rappel : $TSUIC_{pred} = 21.2 - 0.88 \times TEMPT$

Le coefficient de régression associé à TEMPT est de -0.88. Cela signifie que, en moyenne, lorsque la température moyenne d'un pays (TEMPT) augmente d'un degré, le taux de suicide (TSUIC) diminue de **presque 1 pour 100.000** (plus précisément 0.88 pour 100.000).
Là aussi, il faudrait disposer de valeurs repères pour conclure à une liaison faible ou importante entre la température moyenne d'un pays (TEMPT) et le taux de suicide pour 100.000 habitants (TSUIC).

Coefficient de corrélation

```
Menu Statistiques
– Liaisons VI*VD
- VI*VD - Corrélation (r Bravais-Pearson)
```

La corrélation est de -.54.[4] On peut considérer que le lien est fort entre ces deux variables dans la mesure où la corrélation est (en valeur absolue) supérieure à .40.

Dans la population ?

Coefficient de régression?

```
Menu Statistiques
– Régression linéaire
- IC sur coefficient b
```

Tableau 9 : Intervalle de confiance sur le coefficient de régression b

Variables	b	Erreur-ty	IC_inf	IC_sup
TEMPT	-0.88	0.43	-1.83	0.07

Cet intervalle de confiance (*IC* 95% =[-1.83 ; 0.07) nous indique que, dans la population :
- la liaison est peut-être négative, comme dans l'échantillon (entre -1.83 et 0), mais aussi peut-être nulle ou peut-être positive (entre 0 et +0.07). Cet intervalle de confiance nous confirme donc qu'il n'est pas possible de conclure sur le signe de la liaison.

[4] *Le coefficient de régression b et le coefficient de corrélation (r) sont toujours de même signe.*

- on ne peut pas conclure sur la force de la liaison, celle-ci pouvant être nulle mais aussi de presque 2 (-1.83) pour 100.000 habitants que l'on peut considérer comme une valeur importante.
La seule solution est… de recueillir les données pour l'ensemble des pays européens !

Coefficient de corrélation

```
Menu Statistiques
– Liaisons VI*VD
- VI*VD - IC sur corrélation
```

On obtient *IC* 95% = [-.85 ; +.07]. Cet intervalle montre qu'il y a un doute sur le signe de la corrélation parente. On le savait déjà avec le test *t*. Cet intervalle de confiance nous montre que la liaison parente est peut-être forte (comprise entre -.85 et -.40), ou faible (comprise entre -.20 et 0). On ne peut pas conclure sur la force de la liaison parente.

QUALITÉ DE LA PRÉDICTION ?

Le logiciel calculera toujours (sauf très rares exceptions) une équation de régression. La question est alors de savoir si cette équation permet de faire des prédictions précises. Les deux moyens principaux d'évaluer la qualité de la prédiction sont :
- l'analyse des résidus,
- le coefficient de détermination R^2.

Dans l'échantillon ?

Les résidus

```
Menu Statistiques
- Régression linéaire (RL)
- RL - Observé/Prédits/Résidus
```

Le Tableau 10 nous indique, pour chaque pays, le taux de suicide total observé (TSUIC_obs), le TSUIC prédit par la régression (TSUIC_préd) et l'erreur de prédiction appelée résidu (TSUIC_rés).

Tableau 10 : Température, Taux de suicide observé (obs), prédit (préd) et résidus (res) par pays

PAYS	TEMPT	TSUIC_obs	TSUIC_préd	TSUIC_rés
ALLE	8.5	11.38	13.73	-2.35
AUTR	9.1	15.38	13.20	2.18
BELG	9.7	17.30	12.67	4.63
FINL	4.5	21.32	17.24	4.08
NORV	4.2	12.22	17.51	-5.29
ROYA	10.2	6.81	12.23	-5.42
SUED	5.9	11.78	16.01	-4.23
ESPA	14.4	6.49	8.54	-2.05
ITAL	15.7	5.88	7.40	-1.52
PORT	16.0	3.83	7.14	-3.31
HONG	10.9	25.18	11.62	13.56
POLO	7.6	14.24	14.52	-0.28

On constate que ces résidus s'échelonnent de 0.28 à 5.42 pour 100.000 (en valeurs absolues), à l'exception d'un pays atypique (la Hongrie) pour lequel le résidu est de presque 14 pour 100.000. Peuvent-ils être considérés comme faibles ou élevés? Mais pour répondre à cette question, il faudrait disposer de valeurs repères pour qualifier un résidu de faible ou élevé avec ce type de variable (un taux de suicide pour 100.000 habitants).

```
Menu Statistiques
- Régression linéaire (RL)
- RL - Dispersion des résidus
```

Tableau 11 : Dispersion des résidus (minimum, maximum et écart-type)

Variables	Min	Max	Ety
TSUIC_res	-5.4	13.6	5.2

Ce tableau montre le plus grand résidu négatif (Prédit > Observé) et le plus grand résidu positif (Prédit < Observé). Il indique également la moyenne de ces erreurs de prédiction : environ 5 pour 100.000 habitants (ety = 5.2).[5]
Notons que, comme toute moyenne, celle-ci est très influencée (ici tirée vers les valeurs élevées) par une valeur atypique (13.6).

Coefficient de détermination R²

```
Menu Statistiques
– Régression linéaire
– RL - R², R et R²_ajust
```

Le coefficient R^2 :
- est égal à 30%. Cela signifie que TEMPT rend compte de 30% de la variance de TSUIC ou, dit autrement, que TEMPT permet de prédire 30% de la variance de TSUIC.
- sa valeur est élevée (R^2=30% > 16%) on peut conclure que, dans l'échantillon, la prédiction du taux de suicide (TSUIC) par la température (TEMPT) est bonne.

Dans la population ?

Coefficient de détermination?

```
Menu Statistiques
– Régression linéaire
- IC sur R²
```

Tableau 12 : Intervalle de confiance sur R²

	R²	Erreur-ty	IC_inf	IC_sup
TSUIC	29.60%	0.17	0%	67%

On a ici l'intervalle de confiance (*IC* 95%) sur R^2 parent suivant : [0% ; 67%].[6]
Cet *IC* ne fait que confirmer la grande incertitude sur la valeur de R^2 dans l'ensemble des pays européens :
- il est possible que la qualité de la prédiction soit nulle (0%),
- il est possible que la qualité de la prédiction soit bonne (> 16%) voire très bonne (jusqu'à 59% !).

On ne peut donc pas se prononcer, à partir de cet échantillon, sur la possibilité de prédire, par une relation linéaire, le taux de suicide (TSUIC) à partir de la température moyenne (TEMPT) dans l'ensemble des payx européens.
On ne peut pas se prononcer sur la force de cette liaison dans la population. Elle est peut être faible, modérée ou forte (*IC* 95% sur R^2 parent : [0% ; 59%])

[5] *La moyenne arithmétique classique est toujours égale à 0 (la somme des résidus positifs est égale, au signe près, à la somme des résidus négatifs). Cet indice n'apporterait donc aucune information sur l'ampleur des résidus. L'écart type est également une moyenne, mais une "moyenne quadratique". On pourrait également calculer la moyenne des écarts en valeurs absolues.*
[6] *Cette version de SES-Pégase utilise une procédure approchée pour le calcul de l'*IC *(Cohen et al. 2003). Il est possible d'obtenir des valeurs plus précises avec le logiciel SES-Colibri (cf. Menu Outils). On obtient alors :* IC *95% = [0% ; 59%]*

RÉDIGER LE COMPTE RENDU DE L'ANALYSE

On a analysé la liaison, dans certains pays européens, entre deux variables : la température moyenne du pays (TEMPT) et le taux de suicide (TSUIC).
On dispose d'un échantillon de 12 pays européens (cf. Tableau 2),

On constate que,
dans cet échantillon de 12 pays européens en 1999,
la température moyenne
est proche de 10°C et la température médiane est d'environ 9°C.
Toutefois, ces températures varient selon le pays, de 4°C à 16°C
et 50% des pays ont une température moyenne comprise entre 7.7°C et 12.7°C.

Le taux de suicide
est, en moyenne, de 13 pour 100.000 habitants.
Toutefois ce taux varie, d'un pays à l'autre, de 4 à 25 pour 100.000 habitants
et 50% des pays ont un taux de suicide compris entre 6.6 et 16.3 pour 100.000 habitants.

Pour l'échantillon des 12 pays européens, en 1999,
on observe qu'il est possible de prédire, en partie, le taux de suicide à partir de la température moyenne
selon une relation de type linéaire ($TSUIC_{préd}$ = 21.2 - 0.88 x TEMPT).
En moyenne, plus la température du pays augmente, plus le taux de suicide est faible.
Toutefois, pour l'ensemble des pays européens,
on ne peut pas conclure qu'il en est de même
($t_{[10]}$ = 2.05, p = .07 > .05).

Pour l'échantillon des 12 pays européens,
la liaison linéaire entre les deux variables apparaît forte
(r = -.54 < .40).
Toutefois, pour l'ensemble des pays européens,
on ne peut pas conclure sur la force de liaison
(*IC* 95% sur r = [-.85 ; +.07] comprend des valeurs faibles (inférieures à .20 en valeurs absolues) et des valeurs fortes (supérieures à .40 en valeurs absolues).

Pour l'échantillon des 12 pays européens,
la prédiction apparaît de bonne qualité
(R^2 = 30 % > 16 %).
Toutefois, pour l'ensemble des pays européens,
on ne peut pas conclure sur la qualité de la prédiction
(*IC* 95% sur R^2 = [0% ; 59%] comprend des valeurs faibles (< 4%) et des valeurs fortes (> 16%)).

RÉFÉRENCES

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Third Edition)*. Mahwah, New Jersey, London: Lawrence Erlbaum Associates.

Corroyer, D., & Wolff, M. (2003). *L'Analyse Statistique des Données en Psychologie : concepts et méthodes de base*. Paris : Armand Colin, pp. 186-194.

Dodge, Y. (1999). *Analyse de régression appliquée*. Paris: Dunod.

Tenenhaus, M. (2007). *Statistique. Méthodes pour décrire, expliquer et prévoir*. Paris: Dunod.

LISTE DES TABLEAUX

LISTE DES FIGURES

SOMMAIRE