Ce livre blanc fait partie d'une série de documents qui expliquent les recherches menées par les statisticiens de Minitab pour développer les méthodes et les outils de vérification des données utilisés dans l'Assistant de Minitab Statistical Software.

Régression simple

Généralités

La procédure de régression simple de l'Assistant ajuste des modèles linéaire et quadratique comportant un prédicteur continu (X) et une réponse continue (Y) à l'aide de l'estimation des moindres carrés. L'utilisateur peut sélectionner le type de modèle ou permettre à l'Assistant de sélectionner le meilleur modèle d'ajustement. Dans ce document, nous expliquons les critères que l'Assistant utilise pour sélectionner le modèle de régression.

En outre, nous examinons plusieurs éléments importants pour obtenir un modèle de régression valide. Tout d'abord, l'échantillon doit être assez grand pour conférer suffisamment de puissance au test et pour fournir une précision suffisante de l'estimation de l'importance de la relation entre X et Y. Ensuite, il est important d'identifier les données aberrantes susceptibles d'influer sur les résultats de l'analyse. Nous étudions également l'hypothèse selon laquelle le terme d'erreur suit une loi normale et évaluons l'impact de la non-normalité sur les tests d'hypothèse du modèle global et les coefficients. Enfin, pour garantir l'utilité du modèle, il est important que le type de modèle sélectionné reflète précisément la relation entre X et Y.

Sur la base de ces éléments, l'Assistant effectue automatiquement les contrôles suivants sur vos données, et répertorie les résultats dans le rapport :

- Quantité de données
- Données aberrantes
- Normalité
- Ajustement de modèle

Dans ce document, nous étudions l'importance pratique de ces facteurs dans l'analyse de régression et nous décrivons la façon dont nous avons établi notre méthode de contrôle de ces facteurs dans l'Assistant.

Méthodes de régression

Sélection du modèle

L'analyse de régression de l'Assistant est ajustée à un modèle comportant un prédicteur continu et une réponse continue, et peut être ajustée à deux types de modèles :

- Linéaire : $F(x) = \beta_0 + \beta_1 X$
- Quadratique : $F(x) = \beta_0 + \beta_1 X + \beta_2 X^2$

L'utilisateur peut sélectionner le modèle avant d'effectuer l'analyse ou peut autoriser l'Assistant à sélectionner le modèle. Plusieurs méthodes peuvent permettre de déterminer le modèle le plus approprié aux données. Pour garantir l'utilité du modèle, il est important que le type de modèle sélectionné reflète précisément la relation entre X et Y.

Objectif

Nous souhaitions examiner les différentes méthodes susceptibles d'être utilisées pour la sélection du modèle afin de déterminer laquelle utiliser dans l'Assistant.

Méthode

Nous avons examiné trois méthodes généralement utilisées pour la sélection du modèle (Neter et al., 1996). La première méthode identifie le modèle dans lequel le terme d'ordre le plus élevé est significatif. La seconde méthode sélectionne le modèle présentant la valeur R^2_{ajust} maximale. La troisième méthode sélectionne le modèle dans lequel le test F global est significatif. Pour plus de détails, reportez-vous à l'Annexe A.

Pour déterminer l'approche la plus pertinente dans l'Assistant, nous avons examiné les méthodes et comparé leurs calculs. Nous avons également collecté les commentaires d'experts de l'analyse de la qualité.

Les résultats

Sur la base de nos recherches, nous avons décidé d'utiliser la méthode qui sélectionne le modèle en fonction de la signification statistique du terme d'ordre maximal dans le modèle. L'Assistant examine d'abord le modèle quadratique et teste si le terme quadratique (β_2) du modèle est statistiquement significatif. Si ce terme n'est pas significatif, il supprime le terme quadratique du modèle et teste le terme linéaire (β_1). Le modèle sélectionné avec cette approche est présenté dans le rapport de sélection du modèle. En outre, si l'utilisateur a sélectionné un modèle différent de celui sélectionné par l'Assistant, nous l'indiquons dans le rapport de sélection du modèle et dans le rapport.

Nous avons choisi cette méthode en partie sur la base des commentaires de qualiticiens, qui disent privilégier généralement les modèles les plus simples, qui excluent les termes non significatifs. En outre, d'après notre comparaison de méthodes, l'utilisation de la signification statistique du terme le plus élevé du modèle est plus rigoureuse que la méthode qui sélectionne le modèle en fonction de la valeur R^2_{ajust} maximale. Pour plus de détails, reportez-vous à l'Annexe A.

Bien que nous utilisions la signification statistique du terme de modèle le plus élevé pour sélectionner le modèle, nous présentons également la valeur R^2_{ajust} et le test F global pour le modèle du rapport de sélection du modèle. Pour connaître les indicateurs d'état présentés dans le rapport, reportez-vous à la section relative au contrôle de données d'ajustement du modèle ci-dessous.

Contrôles de données

Quantité de données

La puissance fait référence à la probabilité qu'un test d'hypothèse rejette l'hypothèse nulle, lorsqu'elle est fausse. Dans le cas de la régression, l'hypothèse nulle stipule l'absence de relation entre X et Y. Si l'ensemble de données est trop petit, la puissance du test peut ne pas être suffisante pour détecter une relation entre X et Y qui existe réellement. Par conséquent, l'ensemble de données doit être assez grand pour détecter une relation importante d'un point de vue pratique, avec une probabilité élevée.

Objectif

Nous souhaitions déterminer l'impact de la quantité de données sur la puissance du test F global de la relation entre X et Y, ainsi que sur la précision de R^2_{ajust}, qui représente l'estimation de l'importance de la relation entre X et Y. Ces informations sont essentielles. Elles permettent de déterminer si l'ensemble de données est assez grand pour considérer l'importance de la relation observée dans les données comme un indicateur fiable de l'importance sous-jacente réelle de la relation. Pour plus d'informations sur R^2_{ajust}, reportez-vous à l'Annexe A.

Méthode

Pour examiner la puissance du test F global, nous avons effectué des calculs de puissance pour une série d'effectifs d'échantillons et de valeurs R^2_{ajust}. Pour examiner la précision de R^2_{ajust}, nous avons simulé la distribution de R^2_{ajust} pour les différentes valeurs ajustées R^2 (ρ^2_{ajust}) et les différents effectifs d'échantillons. Nous avons examiné la variabilité des valeurs R^2_{ajust} pour déterminer l'effectif d'échantillon recommandé pour que la valeur R^2_{ajust} soit proche de ρ^2_{ajust}. Pour plus d'informations sur les calculs et les simulations, reportez-vous à l'Annexe B.

Les résultats

Nous avons constaté que, en présence d'échantillons relativement grands, la puissance de la régression permet de détecter des relations entre X et Y, même si les relations ne sont pas assez importantes pour revêtir un intérêt pratique. Plus particulièrement, nous avons fait les constats suivants :

- Avec un effectif d'échantillon de 15 et une relation importante entre X et Y (ρ^2_{ajust} = 0,65), la probabilité de détecter une relation linéaire statistiquement significative est de 0,9969. Par conséquent, lorsque le test ne parvient pas à détecter une relation statistiquement significative avec au moins 15 points de données, il est probable que la vraie relation ne soit pas très importante (valeur ρ^2_{ajust} < 0,65).
- Avec un effectif d'échantillon de 40 et une relation relativement faible entre X et Y (ρ^2_{ajust} = 0,25), la probabilité de détecter une relation linéaire statistiquement significative est de 0,9398. Par conséquent, avec 40 points de données, il est probable

que le test F détecte des relations entre X et Y même lorsque la relation est relativement faible.

La régression peut détecter assez facilement des relations entre X et Y. Par conséquent, si vous détectez une relation statistiquement significative, vous devez également évaluer l'importance de la relation à l'aide de R^2_{ajust}. Nous avons découvert que, si l'effectif d'échantillon n'est pas assez grand, R^2_{ajust} n'est pas très fiable et peut varier sensiblement d'un échantillon à l'autre. Toutefois, avec un effectif d'échantillon de 40 minimum, nous avons constaté que les valeurs R^2_{ajust} sont plus stables et plus fiables. Avec un effectif d'échantillon de 40, vous pouvez être sûr à 90 % que la valeur observée de R^2_{ajust} se trouvera à plus ou moins 0,20 de ρ^2_{ajust} quels que soient la valeur réelle et le type de modèle (linéaire ou quadratique). Pour plus d'informations sur les simulations, reportez-vous à l'Annexe B.

Sur la base de ces résultats, l'Assistant affiche les informations suivantes dans le rapport lorsqu'il vérifie la quantité de données :

Etat	Condition
	**Effectif d'échantillon < 40** Votre effectif d'échantillon n'est pas assez grand pour vous permettre d'obtenir une estimation très précise de l'importance de la relation. Les mesures de l'importance de la relation, telles que le R carré et le R carré ajusté, peuvent varier énormément. Pour obtenir une estimation plus précise, utilisez des échantillons plus grands (en général 40 ou plus). **Effectif d'échantillon < 15** Votre échantillon est suffisamment grand pour que vous obteniez une estimation précise de l'importance de la relation.

Données aberrantes

Dans la procédure de régression de l'Assistant, nous définissons les données aberrantes comme des observations ayant des valeurs résiduelles normalisées importantes ou des valeurs à effet de levier importantes. Ces mesures permettent en général d'identifier les données aberrantes dans l'analyse de régression (Neter et al., 1996). Les données aberrantes pouvant avoir une forte influence sur les résultats de l'analyse, il peut être nécessaire de corriger les données pour que l'analyse soit valide. Toutefois, les données aberrantes peuvent également résulter de la variation naturelle du procédé. Par conséquent, il est important d'identifier la cause du comportement aberrant afin de déterminer comment traiter ces points de données.

Objectif

Nous souhaitions déterminer l'importance des valeurs résiduelles normalisées et des valeurs à effet de levier nécessaire pour qu'un point de données aberrant puisse être signalé.

Méthode

Nous avons élaboré nos indications concernant l'identification d'observations aberrantes sur la base de la procédure de régression standard de Minitab (**Stat** > **Régression** > **Régression**).

Les résultats

VALEURS RESIDUELLES NORMALISEES

La valeur résiduelle normalisée est égale à la valeur résiduelle, e_i, divisée par une estimation de son écart type. Habituellement, une observation est considérée comme aberrante si la valeur absolue de la valeur résiduelle normalisée est supérieure à 2. Néanmoins, cette valeur est quelque peu prudente. En général, environ 5 % de la totalité des observations répondent à ce critère du simple fait du hasard (si les erreurs sont distribuées normalement). Par conséquent, il est important d'étudier l'origine du comportement aberrant pour déterminer si une observation est réellement aberrante.

VALEUR A EFFET DE LEVIER

Les valeurs à effet de levier sont uniquement liées à la valeur X d'une observation et ne dépendent pas de la valeur Y. Une observation est déterminée comme aberrante si la valeur à effet de levier est égale à plus de 3 fois le nombre de coefficients de modèle (p), divisée par le nombre d'observations (n). Encore une fois, il s'agit d'une valeur limite fréquemment utilisée, bien que certains manuels utilisent $\frac{2 \times p}{n}$ (Neter et al., 1996).

Si vos données comprennent des points à effet de levier élevés, déterminez s'ils ont une influence excessive sur le type de modèle sélectionné pour ajuster les données. Par exemple, une valeur X extrême unique peut générer une sélection d'un modèle quadratique au lieu d'un modèle linéaire. Vous devez déterminer si la courbure observée dans le modèle quadratique correspond à votre compréhension du procédé. Si ce n'est pas le cas, ajustez un modèle plus simple aux données ou regroupez des données supplémentaires pour étudier de façon plus approfondie le procédé.

Lors du test des données aberrantes, l'Assistant affiche les indicateurs d'état suivants dans le rapport :

Etat	Condition
	Il n'y a aucun point de données aberrant. Ces points pourraient avoir une grande influence sur les résultats.
	Il existe au moins une valeur résiduelle normalisée élevée ou au moins une valeur à effet de levier élevée. Pour identifier les lignes de la feuille de travail, placez votre curseur sur un point ou utilisez la fonction de balayage de Minitab. Les données aberrantes pouvant fortement influer sur les résultats, essayez d'identifier la cause des aberrations. Corrigez les erreurs de mesure ou d'entrée des données. Supprimez les données associées aux causes spéciales et réexécutez l'analyse.

Normalité

Dans le cas de la régression, l'hypothèse générale stipule que les erreurs aléatoires (ε) sont distribuées normalement. L'hypothèse de normalité est importante lors de la réalisation de tests d'hypothèse des estimations des coefficients (β). Heureusement, même lorsque les

erreurs aléatoires ne sont pas distribuées normalement, les résultats de test sont en général fiables lorsque l'échantillon est assez grand.

Objectif

Nous souhaitions déterminer l'effectif d'échantillon nécessaire pour obtenir des résultats fiables avec la loi normale. Nous souhaitions déterminer dans quelle mesure les résultats de test réels correspondaient au seuil de signification cible (alpha ou taux d'erreur de 1ère espèce) pour le test, c'est-à-dire, si le test rejetait incorrectement l'hypothèse nulle plus souvent ou moins souvent que prévu pour les différentes lois non normales.

Méthode

Pour estimer le taux d'erreur de 1ère espèce, nous avons effectué plusieurs simulations à l'aide de lois asymétriques, à queues lourdes et à queues légères qui s'écartent sensiblement de la loi normale. Nous avons effectué des simulations de modèles linéaire et quadratique à l'aide d'un effectif d'échantillon de 15. Nous avons examiné à la fois le test F global et le test du terme d'ordre le plus élevé dans le modèle.

Pour chaque condition, nous avons effectué 10 000 tests. Nous avons généré des données aléatoires afin que pour chaque test, l'hypothèse nulle soit vraie. Nous avons ensuite effectué les tests en utilisant un seuil de signification cible de 0,05. Nous avons compté le nombre de fois, sur 10 000, où les tests avaient rejeté l'hypothèse nulle, puis nous avons comparé cette proportion au seuil de signification cible. Si le test est adapté, les taux d'erreur de 1ère espèce doivent être très proches du seuil de signification cible. Pour plus d'informations sur les simulations, reportez-vous à l'Annexe C.

Les résultats

Aussi bien pour le test F global que pour le test du terme d'ordre le plus élevé du modèle, la probabilité d'obtenir des résultats statistiquement significatifs ne diffère pas énormément pour chacune des lois non normales. Les taux d'erreur de 1ère espèce se trouvent tous entre 0,038 et 0,0529, très près du seuil de signification cible de 0,05.

Etant donné que le test obtient de bons résultats avec des échantillons relativement petits, l'Assistant ne teste pas la normalité des données. En revanche, l'Assistant vérifie l'effectif de l'échantillon et signale les effectifs d'échantillon inférieurs à 15. L'Assistant affiche les indicateurs d'état suivants dans le rapport de la régression :

Etat	Condition
	L'effectif d'échantillon est d'au moins 15. La normalité n'est donc pas un problème.
	L'effectif d'échantillon étant inférieur à 15, la normalité peut être un problème. Vous devez interpréter la valeur de p avec la plus grande vigilance. Pour les échantillons réduits, l'exactitude de la valeur de p est sensible aux erreurs résiduelles non normales.

Ajustement du modèle

Vous pouvez sélectionner le modèle linéaire ou quadratique avant de procéder à l'analyse de régression ou vous pouvez choisir l'Assistant pour sélectionner le modèle. Plusieurs méthodes permettent de sélectionner un modèle approprié.

Objectif

Nous souhaitions examiner les différentes méthodes utilisées pour la sélection du type de modèle afin de déterminer l'approche à adopter dans l'Assistant.

Méthode

Nous avons examiné trois méthodes généralement utilisées pour la sélection du modèle. La première méthode identifie le modèle dans lequel le terme d'ordre le plus élevé est significatif. La seconde méthode sélectionne le modèle présentant la valeur R^2_{ajust} maximale. La troisième méthode sélectionne le modèle dans lequel le test F global est significatif. Pour plus de détails, reportez-vous à l'Annexe A.

Pour déterminer l'approche utilisée dans l'Assistant, nous avons examiné les méthodes et avons comparé leurs calculs. Nous avons également recueilli les commentaires d'experts de l'analyse de la qualité.

Les résultats

Nous avons décidé d'utiliser la méthode qui sélectionne le modèle en fonction de la signification statistique du terme d'ordre le plus élevé. L'Assistant examine tout d'abord le modèle quadratique et teste si le terme quadratique du modèle (β_3) est statistiquement significatif. Si ce terme n'est pas significatif, il teste le terme linéaire (β_1) dans le modèle linéaire. Le modèle sélectionné avec cette approche est présenté dans le rapport de sélection du modèle. En outre, si l'utilisateur a sélectionné un modèle différent de celui sélectionné par l'Assistant, nous l'indiquons dans le rapport de sélection du modèle et dans le rapport. Pour plus d'informations, reportez-vous à la section Méthodes de régression ci-dessus.

Sur la base de nos résultats, le rapport de l'Assistant affiche l'indicateur d'état suivant :

Etat	Condition
	**Si le modèle de l'utilisateur correspond au meilleur modèle d'ajustement de l'Assistant** Vous devez évaluer les données et l'ajustement du modèle par rapport à vos objectifs. Vérifiez les éléments suivants sur la ligne d'ajustement : • L'échantillon couvre l'étendue de valeurs X de façon appropriée. • Le modèle s'ajuste parfaitement à n'importe quelle courbure des données (évitez de sur-ajuster les données). • La droite s'ajuste parfaitement dans toutes les zones concernées. **Si le modèle de l'utilisateur ne correspond pas au meilleur modèle d'ajustement de l'Assistant** Le rapport de sélection du modèle présente un modèle alternatif susceptible d'être plus adapté.

Références

Neter, J., Kutner, M.H., Nachtsheim, C.J. et Wasserman, W. (1996), *Applied linear statistical models*, Chicago : Irwin.

Annexe A : sélection du modèle

Un modèle de régression reliant un prédicteur X à une réponse Y prend la forme suivante :

$Y = f(X) + \varepsilon$

la fonction f(X) représentant la valeur attendue (moyenne) de Y en fonction de X.

L'Assistant propose deux formes possibles de fonction f(X) :

Type de modèle	f(X)
Linéaire	$\beta_0 + \beta_1 X$
Quadratique	$\beta_0 + \beta_1 X + \beta_2 X^2$

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 quadratiques 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 Y_i et la valeur ajustée $\hat{f}(X_i)$ 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.

Pour déterminer la méthode utilisée dans l'Assistant pour sélectionner le type de modèle, nous avons évalué trois options :

- Signification du terme d'ordre le plus élevé du modèle
- Test F global du modèle
- Valeur R^2 ajustée (R^2_{ajust})

Signification du terme d'ordre le plus élevé du modèle

Dans cette approche, l'Assistant commence par le modèle quadratique. L'Assistant teste les hypothèses relatives au terme quadratique dans le modèle quadratique :

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

Si cette hypothèse nulle est rejetée, l'Assistant conclut que le coefficient de terme quadratique est non nul et sélectionne le modèle quadratique. Sinon, l'Assistant teste les hypothèses relatives au modèle linéaire :

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Test F global

Cette méthode teste le modèle global (linéaire ou quadratique). Pour la forme sélectionnée de fonction de régression f(X), elle teste :

$H_0: f(X)\ est\ constant$

$H_1: f(X)\ n'est\ pas\ constant$

R^2 ajusté

Le R^2 (R^2_{ajust}) ajusté mesure le degré de variabilité de la réponse attribué à X par le modèle. Il existe deux moyens fréquents de mesurer l'importance de la relation observée entre X et Y :

$$R^2 = 1 - \frac{SCE}{STC}$$

Et

$$R^2_{adj} = 1 - \frac{SCE/(n-p)}{STC/(n-1)}$$

$$\text{STC} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

STC désigne la somme totale des carrés, qui mesure la variation des réponses autour de leur moyenne globale $\bar{Y}$. La valeur SCE mesure leur variation autour de la fonction de régression f(X). L'ajustement de la valeur R^2_{ajust} correspond au nombre de coefficients (p) dans le modèle complet, ce qui laisse n – p degrés de liberté pour estimer la variance de ε. La valeur R^2 ne diminue jamais lorsque davantage de coefficients sont ajoutés au modèle. Toutefois, du fait de l'ajustement, la valeur R^2_{ajust} peut diminuer lorsque les coefficients supplémentaires n'améliorent pas le modèle. Ainsi, si l'ajout d'un autre terme au modèle n'explique pas la variance supplémentaire dans la réponse, la valeur R^2_{ajust} diminue, indiquant que le terme supplémentaire n'est pas utile. Par conséquent, la mesure ajustée doit être utilisée à des fins de comparaison des modèles linéaire et quadratique.

Relation entre les méthodes de sélection du modèle

Nous souhaitions examiner la relation entre les trois méthodes de sélection du modèle, leur mode de calcul, ainsi que l'impact des unes sur les autres.

Nous avons d'abord examiné la relation entre le calcul du test F global et celui de la valeur R^2_{ajust}. La statistique F pour le test du modèle global peut être exprimée en termes de SCE et de STC, qui sont également utilisées dans le calcul de R^2_{ajust} :

$$F = \frac{(STC - SCE)/(p-1)}{SCE/(n-p)}$$

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{ajust}}{1-R^2_{ajust}}.$$

Les formules ci-dessus montrent que la statistique F est une fonction croissante de R^2_{ajust}. Ainsi, le test rejette l'hypothèse H_0 si et seulement si la valeur R^2_{ajust} est supérieure à une valeur spécifique déterminée par le seuil de signification (α) du test. Pour illustrer ceci, nous avons calculé la valeur R^2_{ajust} minimum nécessaire pour obtenir une signification statistique du modèle quadratique au niveau d'$\alpha = 0{,}05$ pour différents effectifs d'échantillons présentés dans le tableau 1 ci-dessous. Par exemple, avec n = 15, la valeur R^2_{ajust} du modèle doit être égale à au moins 0,291877 pour que le test F global soit statistiquement significatif.

Tableau 1 : Valeur R^2_{ajust} minimale pour un test F global significatif du modèle quadratique au niveau d'$\alpha = 0{,}05$ avec plusieurs effectifs d'échantillons

Effectif de l'échantillon	R^2_{ajust} minimum
4	0,9925
5	0,90
6	0,773799
7	0,66459
8	0,577608
9	0,508796
10	0,453712
11	0,408911
12	0,371895
13	0,340864
14	0,314512
15	0,291877
16	0,272238
17	0,255044
18	0,239872
19	0,226387

Effectif de l'échantillon	R^2_{ajust} minimum
20	0,214326
21	0,203476
22	0,193666
23	0,184752
24	0,176619
25	0,169168
26	0,162318
27	0,155999
28	0,150152
29	0,144726
30	0,139677
31	0,134967
32	0,130564
33	0,126439
34	0,122565
35	0,118922
36	0,115488
37	0,112246
38	0,109182
39	0,106280
40	0,103528
41	0,100914
42	0,098429
43	0,096064
44	0,093809
45	0,091658
46	0,089603
47	0,087637

Effectif de l'échantillon	R^2_{ajust} minimum
48	0,085757
49	0,083955
50	0,082227

Nous avons ensuite examiné la relation entre le test d'hypothèse du terme d'ordre le plus élevé du modèle et R^2_{ajust}. Le test du terme d'ordre le plus élevé, tel que le terme quadratique dans un modèle quadratique, peut être exprimé en termes de sommes des carrés ou de la valeur R^2_{ajust} du modèle complet (par exemple, quadratique) et de la valeur R^2_{ajust} du modèle réduit (par exemple, linéaire) :

$$F = \frac{SCE(Réduit) - SCE(Complet)}{SCE(Complet)/(n-p)}$$

$$= 1 + \frac{(n-p+1)\left(R^2_{ajust}(Complet) - R^2_{ajust}(Réduit)\right)}{1 - R^2_{ajust}(Complet)}.$$

Les formules indiquent que, pour une valeur fixée de $R^2_{ajust}(Réduit)$, la statistique F est une fonction croissante de $R^2_{ajust}(Complet)$. Elles montrent également dans quelle mesure le test dépend de la différence entre les deux valeurs R^2_{ajust}. En particulier, la valeur du modèle complet doit être supérieure à la valeur du modèle réduit pour que vous puissiez obtenir une valeur F assez grande pour être statistiquement significative. Par conséquent, la méthode qui utilise la signification du terme d'ordre le plus élevé pour sélectionner le meilleur modèle est plus rigoureuse que la méthode qui choisit le modèle présentant la valeur R^2_{ajust} maximale. La méthode du terme d'ordre maximal est également compatible avec la préférence de nombreux utilisateurs pour un modèle plus simple. Ainsi, nous avons décidé d'utiliser la signification statistique du terme d'ordre maximal pour sélectionner le modèle de l'Assistant.

Certains utilisateurs ont tendance à privilégier le modèle le mieux ajusté aux données, c'est-à-dire le modèle présentant la valeur R^2_{ajust} maximale. L'Assistant fournit ces valeurs dans le rapport de sélection du modèle et le rapport.

Annexe B : quantité de données

Dans cette section, nous examinons l'influence de la valeur n, le nombre d'observations, sur la puissance du test de modèle global et sur la précision de R^2_{ajust}, l'estimation de l'importance du modèle.

Pour quantifier l'importance de la relation, nous introduisons une nouvelle quantité, ρ^2_{ajust}, en tant qu'équivalent de la population de la statistique d'échantillon R^2_{ajust}. Souvenez-vous que

$$R^2_{adj} = 1 - \frac{SCE/(n-p)}{STC/(n-1)}$$

Nous définissons donc

$$\rho^2_{adj} = 1 - \frac{E(SCE|X)/(n-p)}{E(STC|X)/(n-1)}$$

L'opérateur E(·|X) indique la valeur attendue, ou la moyenne d'une variable aléatoire en fonction de la valeur de X. En supposant que le modèle correct est $Y = f(X) + \varepsilon$ avec une valeur ε indépendante distribuée de façon identique, nous avons

$$\frac{E(SCE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$

$$\frac{E(STC|X)}{n-1} = \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1)+\sigma^2} + \sigma^2 \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1)+\sigma^2}$$

$\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

D'où,

$$\rho^2_{ajust} = \frac{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1)}{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1) + \sigma^2}$$

Signification du modèle global

Lorsque nous testons la signification statistique du modèle global, nous supposons que les erreurs aléatoires ε sont indépendantes et distribuées normalement. Ensuite, dans l'hypothèse nulle où la valeur de la moyenne Y est constante ($f(X) = \beta_0$), la statistique du test F a une distribution $F(p-1, n-p)$. Dans l'hypothèse alternative, la statistique F a une distribution $F(p-1, n-p, \theta)$ non centrée avec un paramètre de non-centralité :

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho^2_{ajust}}{1-\rho^2_{ajust}}$$

La probabilité de rejeter l'hypothèse H0 augmente avec le paramètre de non-centralité, dont les valeurs n et ρ^2_{ajust}. augmentent.

A l'aide de la formule ci-dessus, nous avons calculé la puissance des tests F globaux pour une série de valeurs ρ^2_{ajust} lorsque n = 15 pour les modèles linéaire et quadratique. Les résultats sont présentés dans le tableau 2.

Tableau 2 : Puissance des modèles linéaire et quadratique avec différentes valeurs ρ^2_{ajust} et n = 15

ρ^2_{ajust}	θ	Puissance de F Linéaire	Puissance de F Quadratique
0,05	**0,737**	0,12523	0,09615
0,10	**1,556**	0,21175	0,15239
0,15	**2,471**	0,30766	0,21896
0,20	**3,50**	0,41024	0,2956
0,25	**4,667**	0,5159	0,38139
0,30	**6,00**	0,62033	0,47448
0,35	**7,538**	0,71868	0,57196
0,40	**9,333**	0,80606	0,66973
0,45	**11,455**	0,87819	0,76259
0,50	**14,00**	0,93237	0,84476
0,55	**17,111**	0,96823	0,91084
0,60	**21,00**	0,9882	0,95737
0,65	**26,00**	0,99688	0,98443
0,70	**32,667**	0,99951	0,99625
0,75	**42,00**	0,99997	0,99954
0,80	**56,00**	1,00	0,99998
0,85	**79,333**	1,00	1,00
0,90	**126,00**	1,00	1,00
0,95	**266,00**	1,00	1,00

Dans l'ensemble, nous avons établi que le test possède une puissance élevée lorsque la relation entre X et Y est importante et que l'effectif d'échantillon est d'au moins 15. Par exemple, lorsque la valeur ρ^2_{ajust} = 0,65, le tableau 2 montre que la probabilité de détecter une relation linéaire statistiquement significative au niveau d'$\alpha = 0{,}05$ est de 0,99688. L'incapacité à détecter une relation aussi importante avec le test F concernerait moins de 0,5 % des échantillons. Même pour un modèle quadratique, l'incapacité à détecter la relation

avec le test F concernerait moins de 2 % des échantillons. Par conséquent, lorsque le test ne parvient pas à détecter une relation statistiquement significative avec au moins 15 observations, il est intéressant de noter que la vraie relation, s'il y en a bien une, a une valeur de ρ^2_{ajust} inférieure à 0,65. Notez qu'il n'est pas nécessaire que la valeur ρ^2_{ajust} soit de 0,65 pour avoir un intérêt pratique.

Nous souhaitions également examiner la puissance du test F global lorsque l'effectif d'échantillon était supérieur (n = 40). Nous avons déterminé que l'effectif d'échantillon n = 40 est un seuil important pour la précision de la valeur R^2_{ajust} (voir la section Importance de la relation ci-dessous) et nous souhaitions évaluer les valeurs de puissance de l'effectif d'échantillon. Nous avons calculé la puissance des tests F globaux pour une série de valeurs ρ^2_{ajust} lorsque n = 40 pour les modèles linéaire et quadratique. Les résultats sont présentés dans le tableau 3.

Tableau 3 : Puissance des modèles linéaire et quadratique avec différentes valeurs ρ^2_{ajust} et n = 40

ρ^2_{ajust}	θ	Puissance de F Linéaire	Puissance de F Quadratique
0,05	**2,0526**	0,28698	0,21541
0,10	**4,3333**	0,52752	0,41502
0,15	**6,8824**	0,72464	0,60957
0,20	**9,75**	0,86053	0,76981
0,25	**13,00**	0,9398	0,88237
0,30	**16,7143**	0,97846	0,94925
0,35	**21,00**	0,99386	0,98217
0,40	**26,00**	0,99868	0,99515
0,45	**31,9091**	0,9998	0,99905
0,50	**39,00**	0,99998	0,99988
0,55	**47,6667**	1,00	0,99999
0,60	**58,50**	1,00	1,00
0,65	**72,4286**	1,00	1,00

Nous avons trouvé que la puissance était élevée, même lorsque la relation entre X et Y était relativement faible. Par exemple, même lorsque ρ^2_{ajust} = 0,25, le tableau 3 montre que la probabilité de détecter une relation linéaire statistiquement significative au niveau d'$\alpha =$ 0,05 est de 0,93980. Avec 40 observations, il est peu probable que le test F ne parvienne pas à détecter une relation entre X et Y, même si cette relation est relativement faible.

Importance de la relation

Comme nous l'avons déjà démontré, une relation significative sur le plan statistique au niveau des données n'indique pas nécessairement une forte relation sous-jacente entre X et Y. C'est pourquoi de nombreux utilisateurs comptent sur des indicateurs tels que R^2_{ajust} pour connaître l'importance réelle de la relation. Si nous considérons R^2_{ajust} comme une estimation de ρ^2_{ajust}, nous voulons avoir la certitude que l'estimation est relativement proche de la valeur ρ^2_{ajust} réelle.

Pour illustrer la relation entre R^2_{ajust} et ρ^2_{ajust}, nous avons simulé la distribution de R^2_{ajust} pour les différentes valeurs de ρ^2_{ajust} afin de connaître la variabilité de R^2_{ajust} pour les différentes valeurs de n. Les graphiques des figures 1-4 ci-dessous sont des histogrammes de 10 000 valeurs simulées de R^2_{ajust}. Dans chaque paire d'histogrammes, la valeur de ρ^2_{ajust} est la même, ce qui nous permet de comparer la variabilité de R^2_{ajust} pour les échantillons d'effectif 15 et les échantillons d'effectif 40. Nous avons testé les valeurs ρ^2_{ajust} de 0,0, 0,30, 0,60 et 0,90. Toutes les simulations ont été effectuées avec le modèle linéaire.

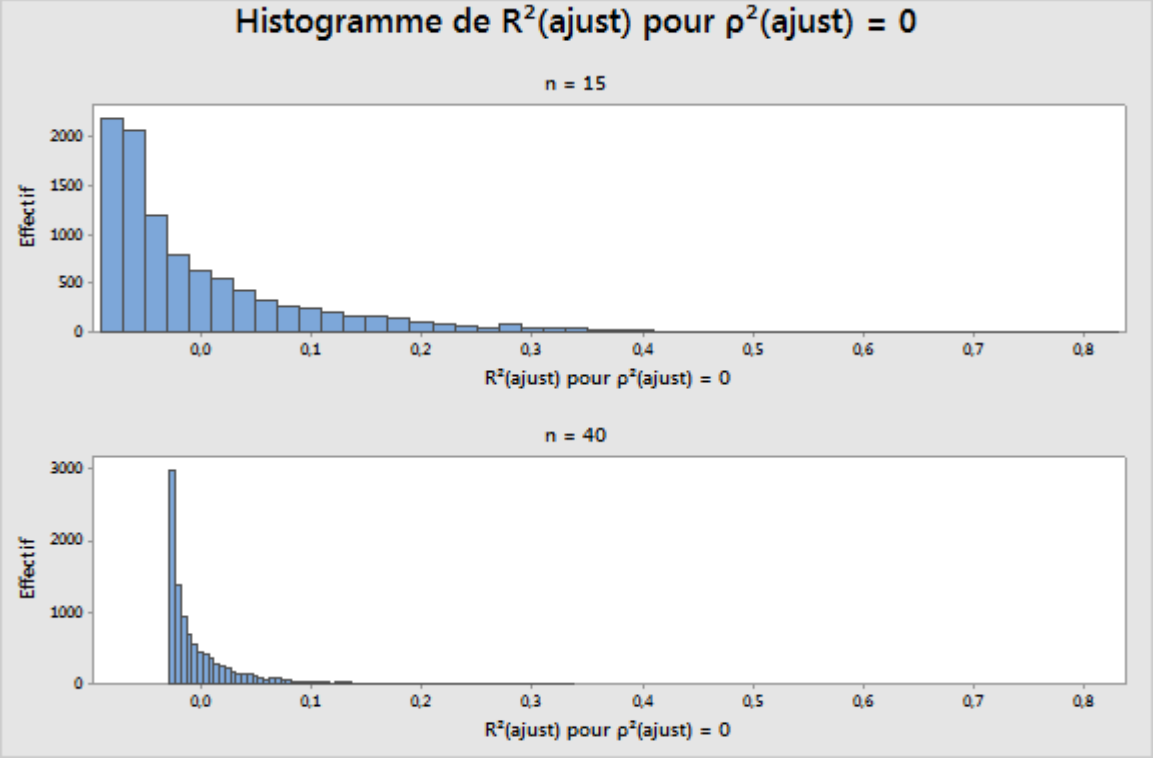

Figure 1 : Valeurs R^2_{ajust} simulées pour ρ^2_{ajust} = 0,0 pour n = 15 et n = 40

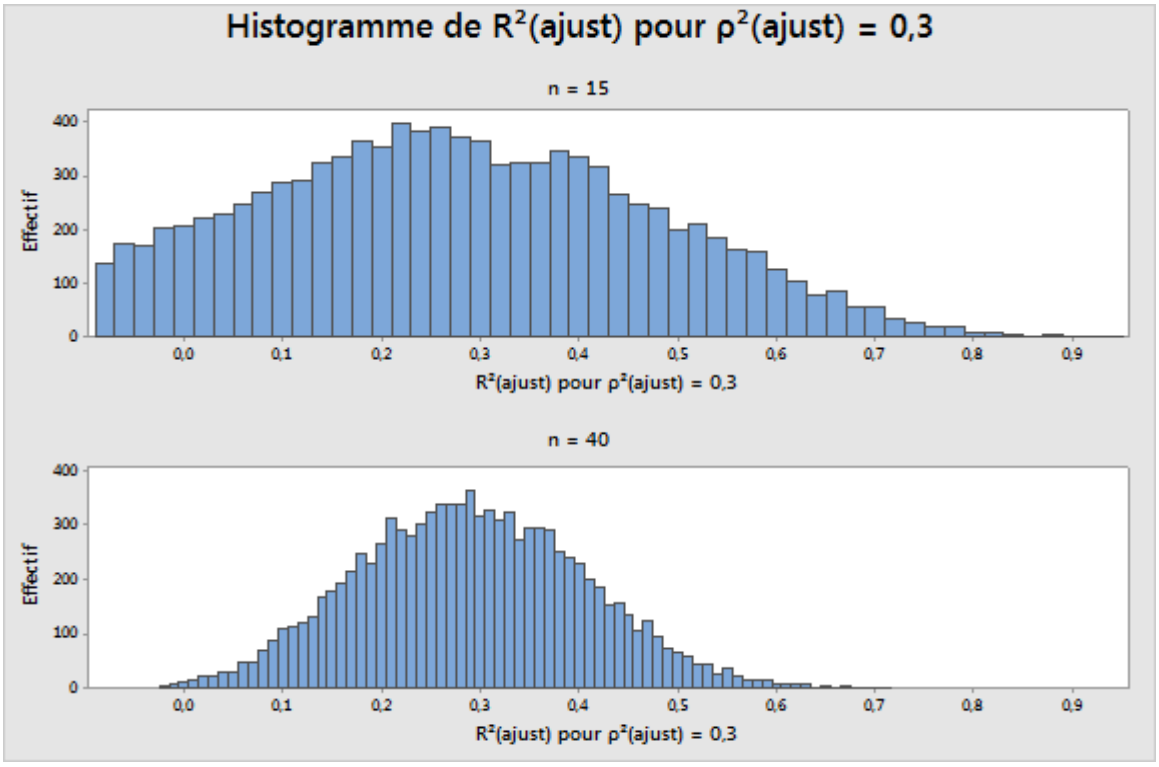

Figure 2 : Valeurs R^2_{ajust} simulées pour ρ^2_{ajust} = 0,30 pour n = 15 et n = 40

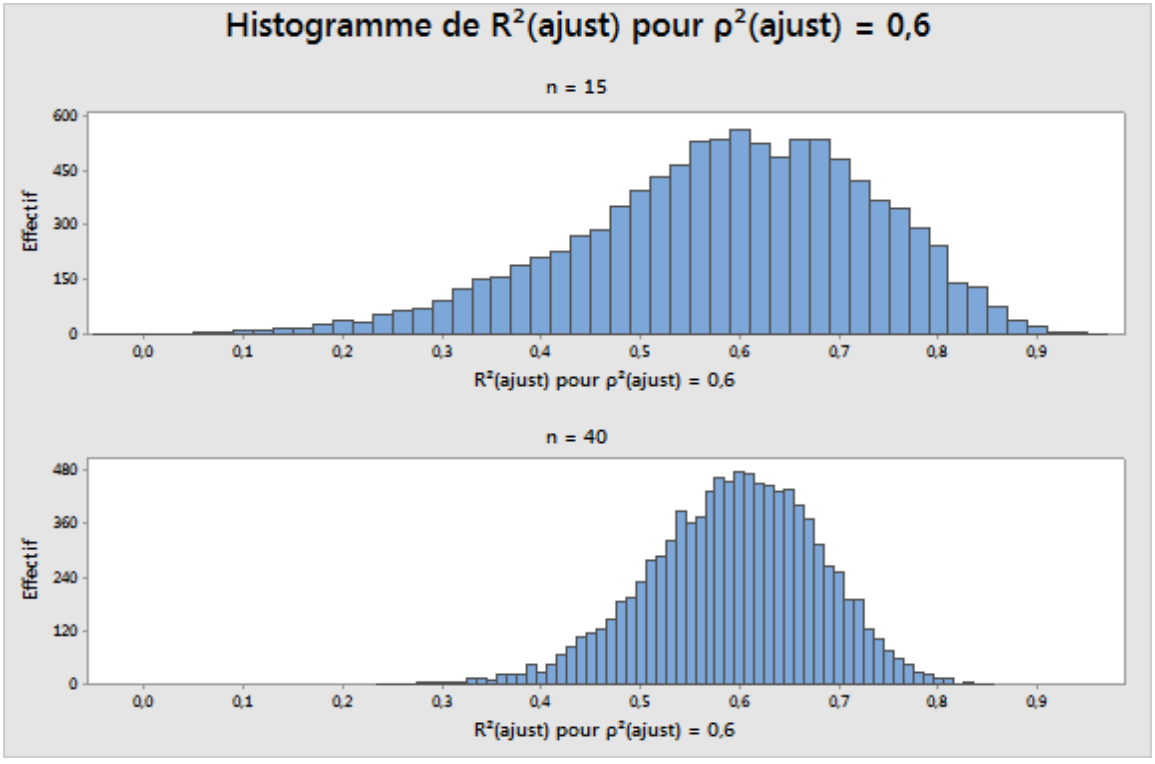

Figure 3 : Valeurs R^2_{ajust} simulées pour ρ^2_{ajust} = 0,60 pour n = 15 et n = 40

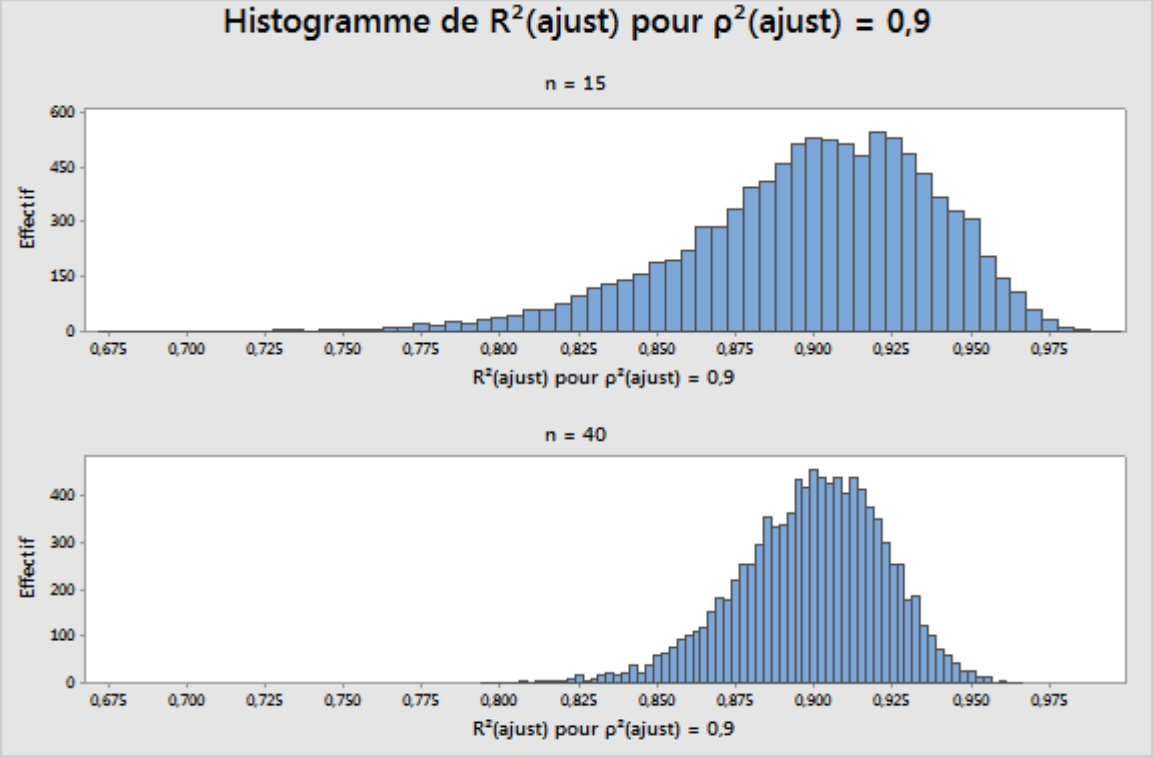

Figure 4 : Valeurs R^2_{ajust} simulées pour ρ^2_{ajust} = 0,90 pour n = 15 et n = 40

Dans l'ensemble, les simulations montrent qu'il peut y avoir une différence considérable entre l'importance réelle de la relation (ρ^2_{ajust}) et la relation observée dans les données (R^2_{ajust}). Augmenter l'effectif d'échantillon de 15 à 40 réduit considérablement l'importance probable de la différence. Nous avons établi que 40 observations constituent un seuil approprié en identifiant la valeur minimale de n pour laquelle des différences absolues $|R^2_{ajust} - \rho^2_{ajust}|$ supérieures à 0,20 surviennent avec une probabilité maximale de 10 %. Cette conclusion est indépendante de la valeur réelle de ρ^2_{ajust} dans chacun des modèles considérés. Pour le modèle linéaire, le cas le plus difficile était ρ^2_{ajust} = 0,31, qui nécessitait une valeur n = 36. Pour le modèle quadratique, le cas le plus difficile était ρ^2_{ajust} = 0,30, qui nécessitait la valeur n = 38. Avec 40 observations, vous pouvez être sûr à 90 % que la valeur observée de R^2_{ajust} se situera à plus ou moins 0,20 de ρ^2_{ajust}, indépendamment de cette valeur et que vous utilisiez le modèle linéaire ou le modèle quadratique.

Annexe C : normalité

Les modèles de régression utilisés dans l'Assistant présentent tous la même forme :

$$Y = f(X) + \varepsilon$$

L'hypothèse générale autour des termes aléatoires ε stipule que ce sont des variables aléatoires normales indépendantes et distribuées de façon identique avec une moyenne nulle et une variance commune σ^2. Les estimations par la méthode des moindres carrés des paramètres β constituent tout de même les meilleures estimations linéaires non biaisées, même si nous renonçons à l'hypothèse selon laquelle les valeurs ε sont distribuées normalement. L'hypothèse de normalité devient seulement importante lorsque nous tentons de lier les probabilités à ces estimations, comme nous le faisons dans les tests d'hypothèse autour de f(X).

Nous souhaitions déterminer la valeur n nécessaire pour que les résultats d'une analyse de régression en fonction de l'hypothèse de normalité soient fiables. Nous avons procédé à des simulations pour comparer les taux d'erreur de 1ère espèce des tests d'hypothèse avec différents lois de distribution non normales de l'erreur.

Le tableau 4 ci-dessous montre la proportion de simulations, parmi 10 000, dans lesquelles le test F global était significatif au niveau d'$\alpha = 0{,}05$ pour plusieurs distributions de la valeur ε pour les modèles linéaire et quadratique. Dans ces simulations, l'hypothèse nulle, qui indique l'absence de relation entre X et Y, était vraie. Les valeurs X étaient espacées de façon homogène sur un intervalle. Nous avons utilisé un effectif d'échantillon de n = 15 pour tous les tests.

Tableau 4 : Taux d'erreur de 1ère espèce pour les tests F globaux pour les modèles linéaire et quadratique avec n = 15 pour les lois non normales

Loi de distribution	Linéaire significatif	Quadratique significatif
Normale	0,04770	0,05060
t(3)	0,04670	0,05150
t(5)	0,04980	0,04540
Laplace	0,04800	0,04720
Uniforme	0,05140	0,04450
Beta(3, 3)	0,05100	0,05090
Exponentielle	0,04380	0,04880
Chi(3)	0,04860	0,05210
Chi(5)	0,04900	0,05260
Chi(10)	0,04970	0,05000

Loi de distribution	Linéaire significatif	Quadratique significatif
Bêta(8, 1)	0,04780	0,04710

Nous avons ensuite examiné le test du terme d'ordre le plus élevé utilisé pour sélectionner le meilleur modèle. Pour chaque simulation, nous avons étudié si le terme quadratique était significatif. Dans les cas où le terme quadratique ne l'était pas, nous avons étudié si le terme linéaire était significatif. Dans ces simulations, l'hypothèse nulle était vraie, pour un niveau cible d'$\alpha = 0{,}05$ et n = 15.

Tableau 5 : Taux d'erreur de 1ère espèce pour les tests du terme d'ordre le plus élevé pour les modèles linéaire ou quadratique avec n = 15 pour les lois non normales

Loi de distribution	Quadratique	Linéaire
Normale	0,05050	0,04630
t(3)	0,05120	0,04300
t(5)	0,04710	0,04820
Laplace	0,04770	0,04660
Uniforme	0,04670	0,04900
Beta(3, 3)	0,05000	0,04860
Exponentielle	0,04600	0,03800
Chi(3)	0,05110	0,04290
Chi(5)	0,05290	0,04490
Chi(10)	0,04970	0,04610
Bêta(8, 1)	0,04770	0,04380

Les résultats de la simulation montrent que, pour le test F global et pour le test du terme d'ordre le plus élevé du modèle, la probabilité d'obtenir des résultats statistiquement significatifs ne diffère pas considérablement pour chacune des distributions d'erreurs. Les taux d'erreur de 1ère espèce se trouvent tous entre 0,038 et 0,0529.