Au secours! Mon modèle passe en Prod … JMPy à la rescousse

Aide à la régression univariée – Cas de 2 variables X et Y continues
Etude de cas détaillée

JMP Addict – Atelier Trucs et Astuces – Webinar du 9 Juin 2022

Etape 1
Analyse exploratoire des données

1a. Graphique

Un graphique vaut mieux qu'un long discours! Représente Y en fonction de X.

Avant toute modélisation, il est important d'explorer les données (min, max, range) et de visualiser le comportement de Y en fonction de X au moyen d'un graphique.

Plateformes JMP

1. Via le constructeur de graphique (Graph Builder)
2. Via la plateforme Ajuster Y en fonction de X (Fit X by Y)

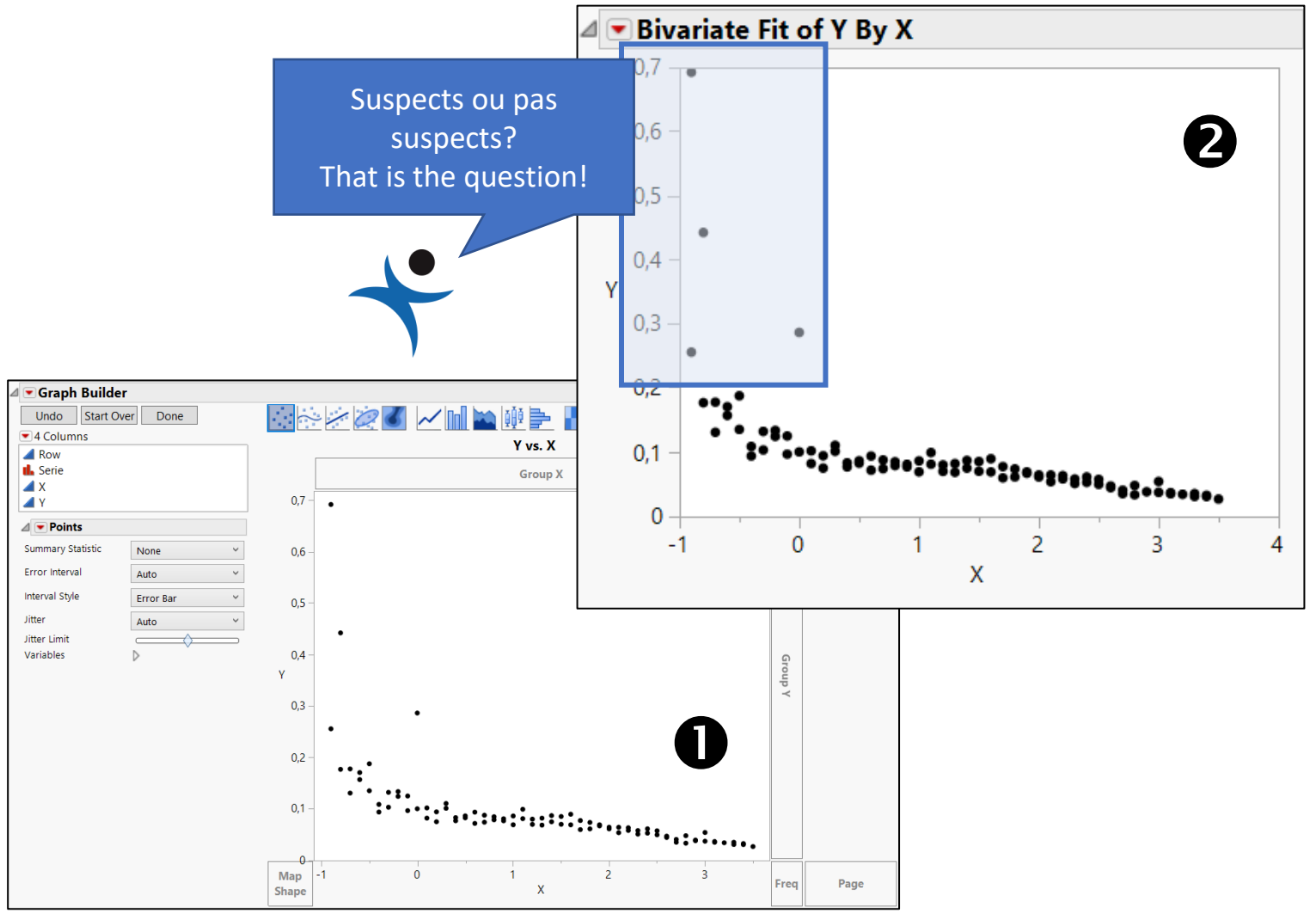
Les conseils de JMPy

Mène les premières observations.
Instinctivement, quel type de modélisation penses-tu utiliser? Linéaire? Non-linéaire? Des données te paraissent-elles suspectes? Ce n'est pas évident, n'est-ce pas? Plusieurs cas sont possibles:

- Les valeurs de Y>0.2 sont suspectes? Dans ce cas, un modèle linéaire conviendrait peut-être.
- Les valeurs de Y>0.2 sont plausibles? Dans ce cas, le comportement semble non linéaire

Avançons prudemment pas à pas.

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 1
Analyse exploratoire des données

1b. Suspects?

Des points sont-ils suspects? Peut-être doivent-ils être écartés pour le moment?

Au premier coups d'œil, existe-t-il des points suspects qui mériteraient d'être écartés temporairement? … écartés! Pas supprimés! On reviendra sur eux plus tard.

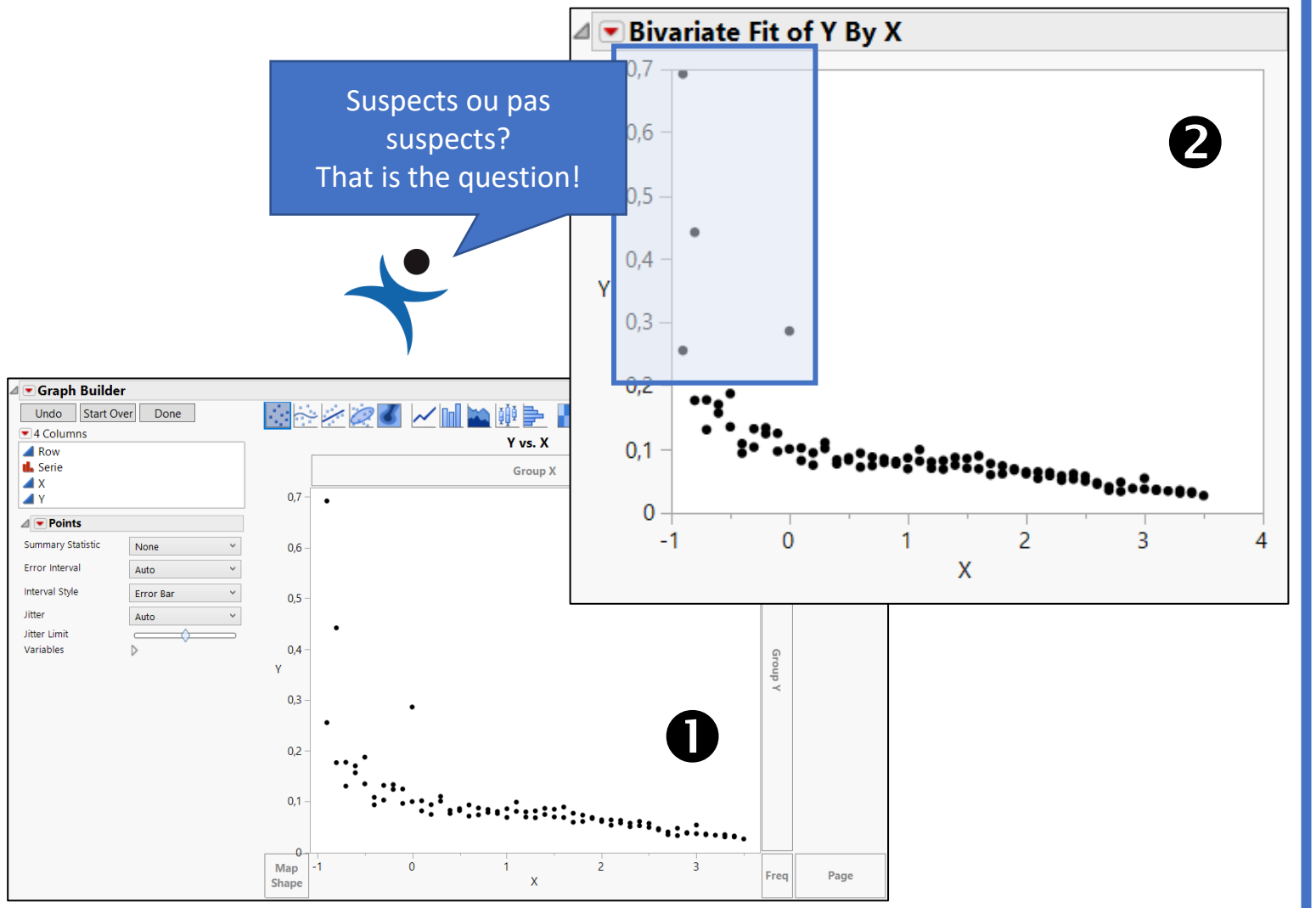
Plateforme JMP

1. Via le constructeur de graphique (Graph Builder)
2. Via la plateforme Ajuster Y en fonction de X (Fit X by Y)

Les conseils de JMPy

Dans le doute, mets de côté les points qui te paraissent suspects (Y>0.2). Ne les efface pas, on reviendra sur eux plus tard

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 1
Analyse exploratoire des données

1c. Corrélation

Combien vaut le coefficient de corrélation? Pas la peine d'aller plus loin si X et Y ne sont pas liés entre eux!

Existe-t-il un lien entre X et Y? L'analyse du coefficient de corrélation de Pearson (noté r) va permettre de quantifier la force de la relation linéaire entre X et Y.▼

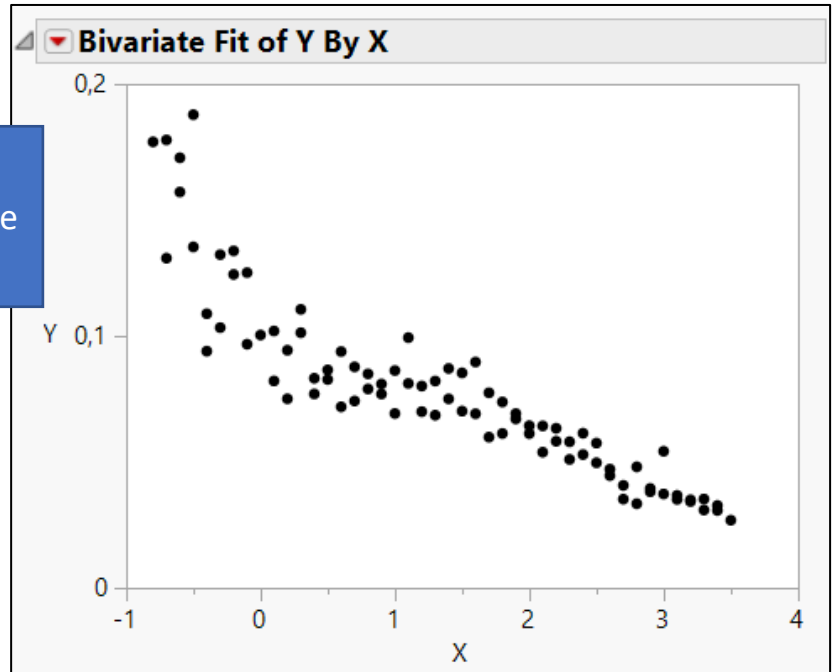
Plateforme JMP

Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Statistiques de résumé (Summary Statistics)

Les conseils de JMPy

Le coefficient de corrélation r de Pearson est négatif et vaut -0.61533, significatif avec une p-value très faible. Il indique que la relation entre X et Y est forte.

La relation entre X et Y semble être forte, ça vaut le coup de poursuivre.



Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0,61533	-0,73003	-0,4669	<,0001*
Covariance	-0,06946			
Count	89			

Variable	Mean	Std Dev
X	1,275281	1,292046
Y	0,093179	0,087366

Envie de rejoindre la communauté des heureux utilisateurs de JMP?

Etape 2
Hypothèse d'un modèle linéaire

2a. Regression

Ne sors pas l'artillerie lourde toute de suite! Trace la droite de régression linéaire

Comme il semble qu'il y ait une corrélation entre X et Y, il est temps de tracer la droite de régression, i.e. la droite que l'on peut tracer dans le nuage de points qui représente le mieux la distribution à 2 caractères étudiées.

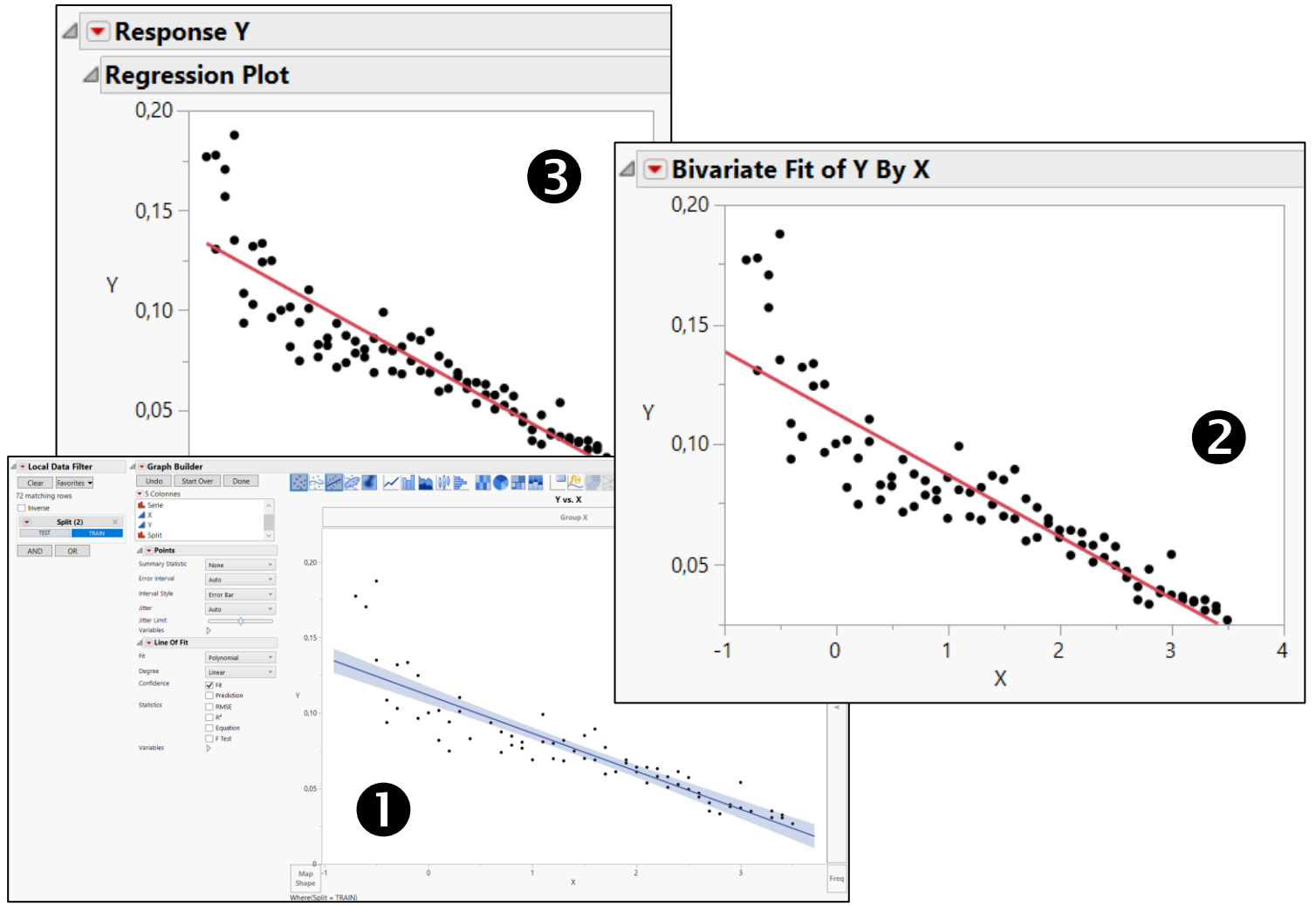
Plateformes JMP

1. Via le constructeur de graphique (Graph Builder), sélectionne Droite d'ajustement (Line of fit)
2. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Regression simple (Fit Line)
3. Via la plateforme Modèle Linéaire (Fit Model), sélectionne Rapport Minimal en emphase.

Les conseils de JMPy

-

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 2
Hypothèse d'un modèle linéaire

2c. Intérêt

Le modèle a-t-il un intérêt? Est-il meilleur que la simple moyenne?

Notre modèle a-t-il finalement un intérêt? Est-il meilleur que la simple moyenne? Nous allons vérifier cela ensemble.

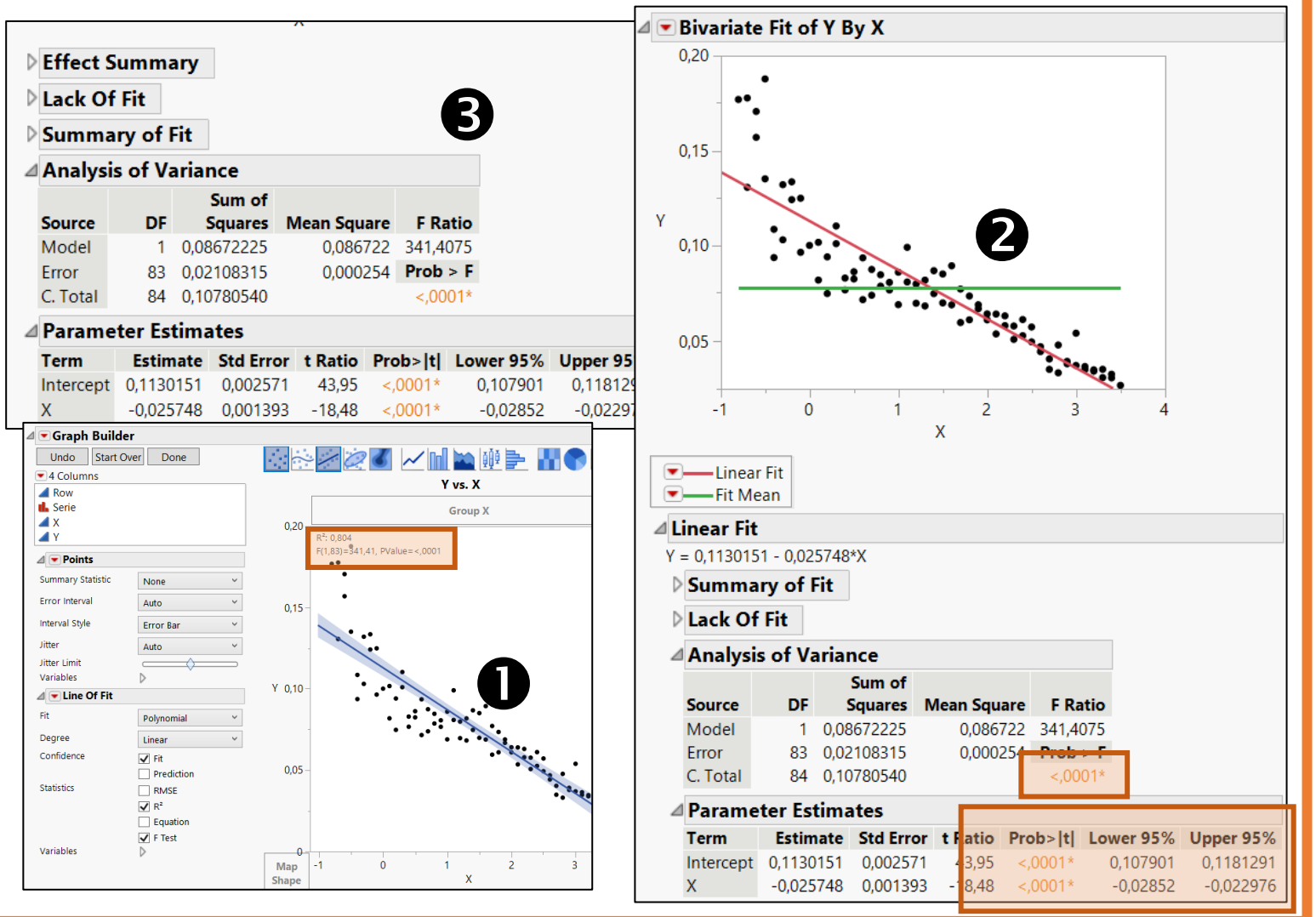
Plateformes JMP

1. Via le constructeur de graphique (Graph Builder). Cocher F-Test (test de Fisher)
2. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Ajustement à la moyenne (Fit Mean) pour afficher la moyenne (pour information)
3. Via la plateforme Modèle Linéaire (Fit Model). Sélectionner Rapport Minimal en emphase.

Les conseils de JMPy

1) La p-value du test de l'analyse de la variance est très faible, la régression est meilleure que la simple moyenne. 2) Les p-values de l'estimation des paramètres de régression sont faibles indiquant que les paramètres sont significativement différents de 0. 3) L'amplitude de l'erreur standard est faible (Plus elle est faible, meilleur est la qualité du modèle). Un paramètre sera jugé non significatif si la valeur 0 appartient à l'intervalle [Lower95%, Upper95%].

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 2
Hypothèse d'un modèle linéaire

2b. R^2

Le coefficient de détermination est-il satisfaisant pour ton cas d'application?

Le pouvoir explicatif (coefficient de détermination R^2, correspondant à la part de variabilité expliquée par le modèle de régression) est-il satisfaisant en fonction du cas d'usage?

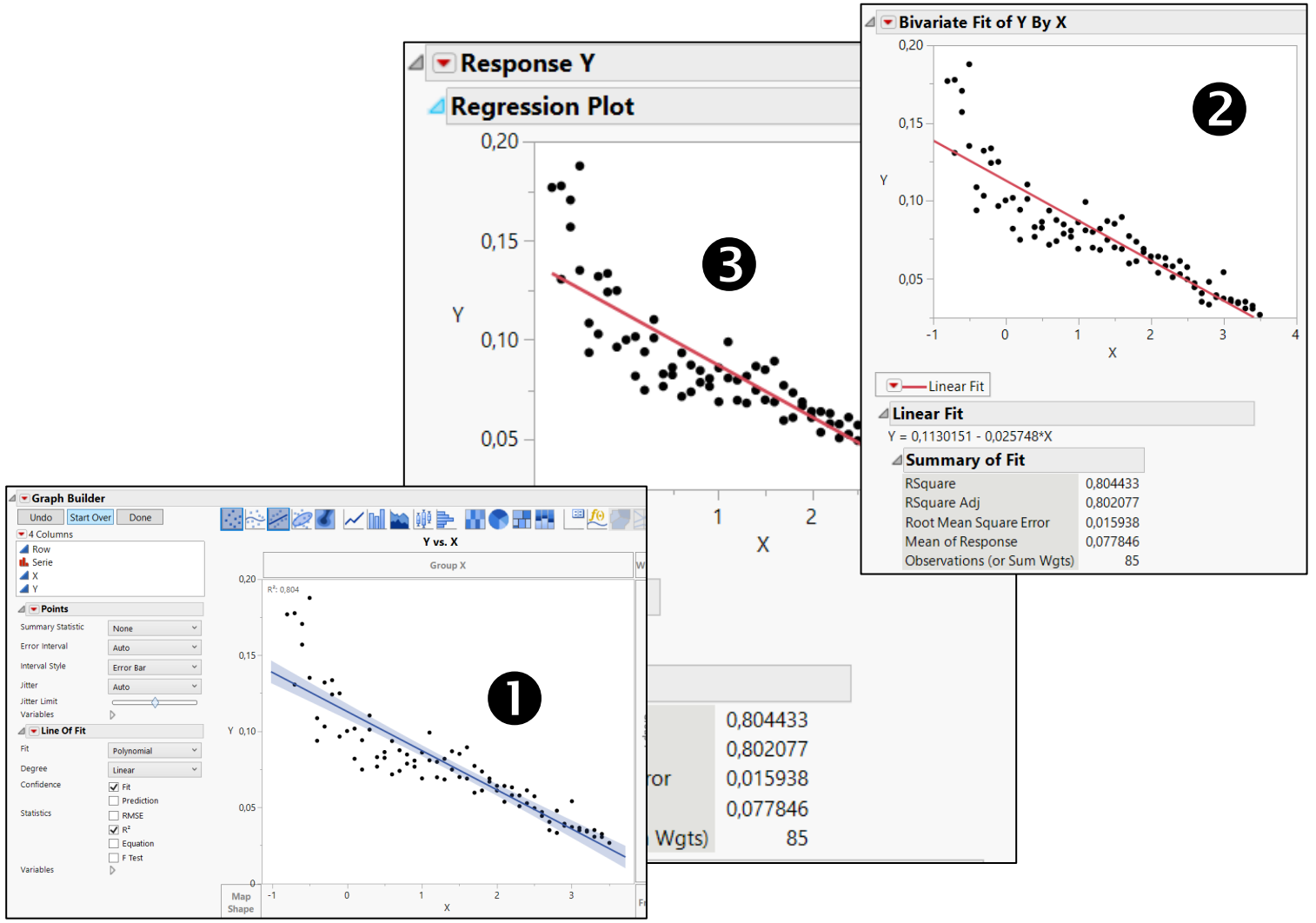
Plateformes JMP

1. Via le constructeur de graphique (Graph Builder). Cocher R^2, Equation
2. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Regression simple (Fit Line)
3. Via la plateforme Modèle Linéaire (Fit Model). Sélectionner Rapport Minimal en emphase.

Les conseils de JMPy

Le coefficient de détermination R^2 vaut 0.804433, il permet de juger la qualité d'une régression linéaire simple. Aussi appelé pouvoir explicatif, il traduit dans le cas présent que la simple variable X permet d'expliquer 80.4% de toute la variabilité observée. Est-il satisfaisant? Seul toi peut répondre (parfois, 50% est jugé suffisant!).

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 3
Validation du modèle

3a. Lack of Fit

Tu as des valeurs répétées? Vérifie s'il n'y a pas un défaut d'ajustement du modèle?

Dans le cas où l'on dispose de valeurs répétées, il est possible de détecter s'il y a un manque d'adéquation du modèle - aussi appelé défaut d'ajustement (Lack of Fit)

Plateformes JMP

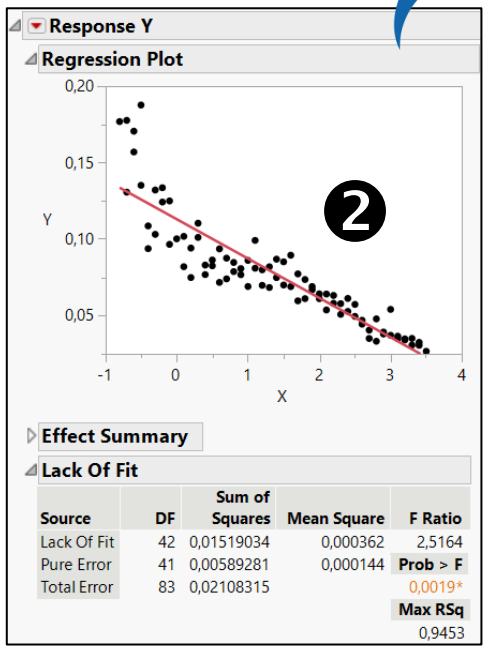
1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), Section « Défaut d'ajustement » (Lack of Fit)
2. Via la plateforme Modèle Linéaire (Fit Model).

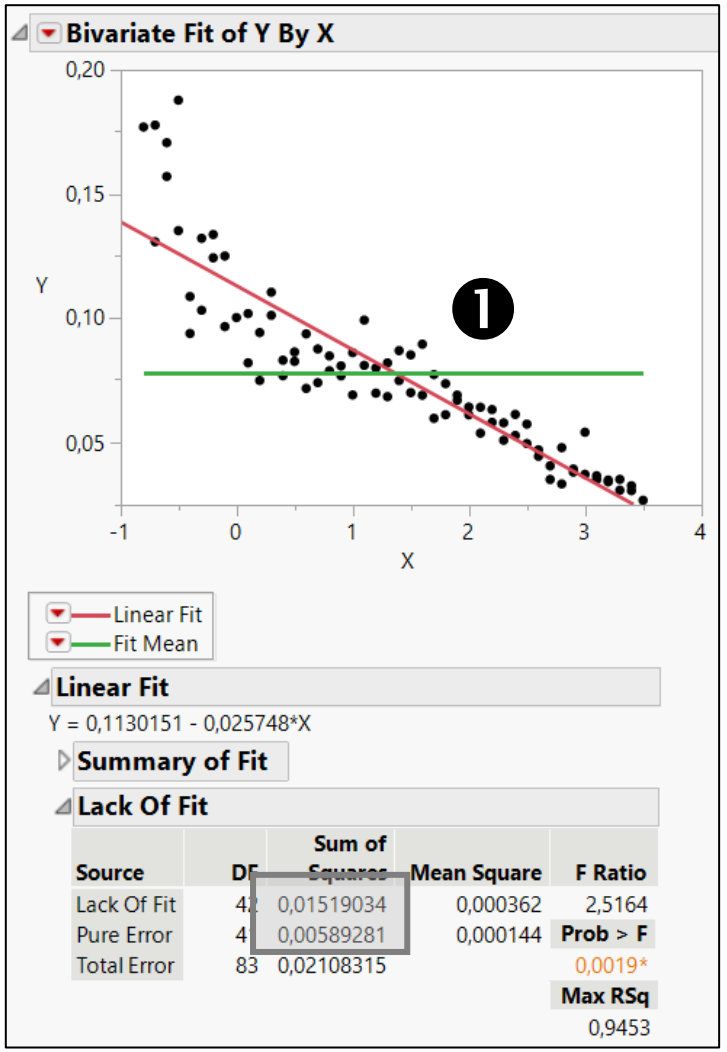
Les conseils de JMPy

Une erreur de manque d'ajustement (Lack Of Fit) significativement plus grande que l'erreur pure (Pure Error) indique qu'il reste quelque chose dans les résidus qui peut être éliminé par un modèle plus approprié. Dans notre cas, l'erreur de manque d'ajustement est 2-3 fois plus grande que celle de l'erreur pure, la p-value est significative, ce qui constitue une indication que des termes importants ont été oubliées, comme par exemple des termes quadratiques.

Envie de rejoindre la communauté des heureux utilisateurs de JMP?

Qu'est-ce que l'absence d'ajustement pour une régression linéaire? Un modèle de régression présente un manque d'ajustement lorsqu'il ne parvient pas à décrire de manière adéquate la relation fonctionnelle entre les facteurs expérimentaux et la variable de réponse. Un manque d'ajustement peut se produire si des termes importants du modèle (termes quadratiques, autres facteurs), ne sont pas inclus.





Etape 3
Validation du modèle

3b. Résidus

Sont-ils:
- Homoscédastiques?
- Approx. norm. distribués?
- Indépendants?

L'analyse des résidus permet de s'assurer qu'il ne reste plus aucune information à extraire des résidus, que toutes les tendances sont bien incorporées dans le modèle et que seul le bruit de fond (aléatoire) demeure.

Plateformes JMP

1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Régression simple (Linear Fit), puis sélectionne Tracer les résidus (Plot Residuals)
2. Via la plateforme Modèle Linéaire (Fit Model). ▼ Réponse (Response), puis Diagnostics de lignes (Row Diagnostics) > Tracer les résidus en fonction des valeurs prévues (Residual by Predicted Plot)

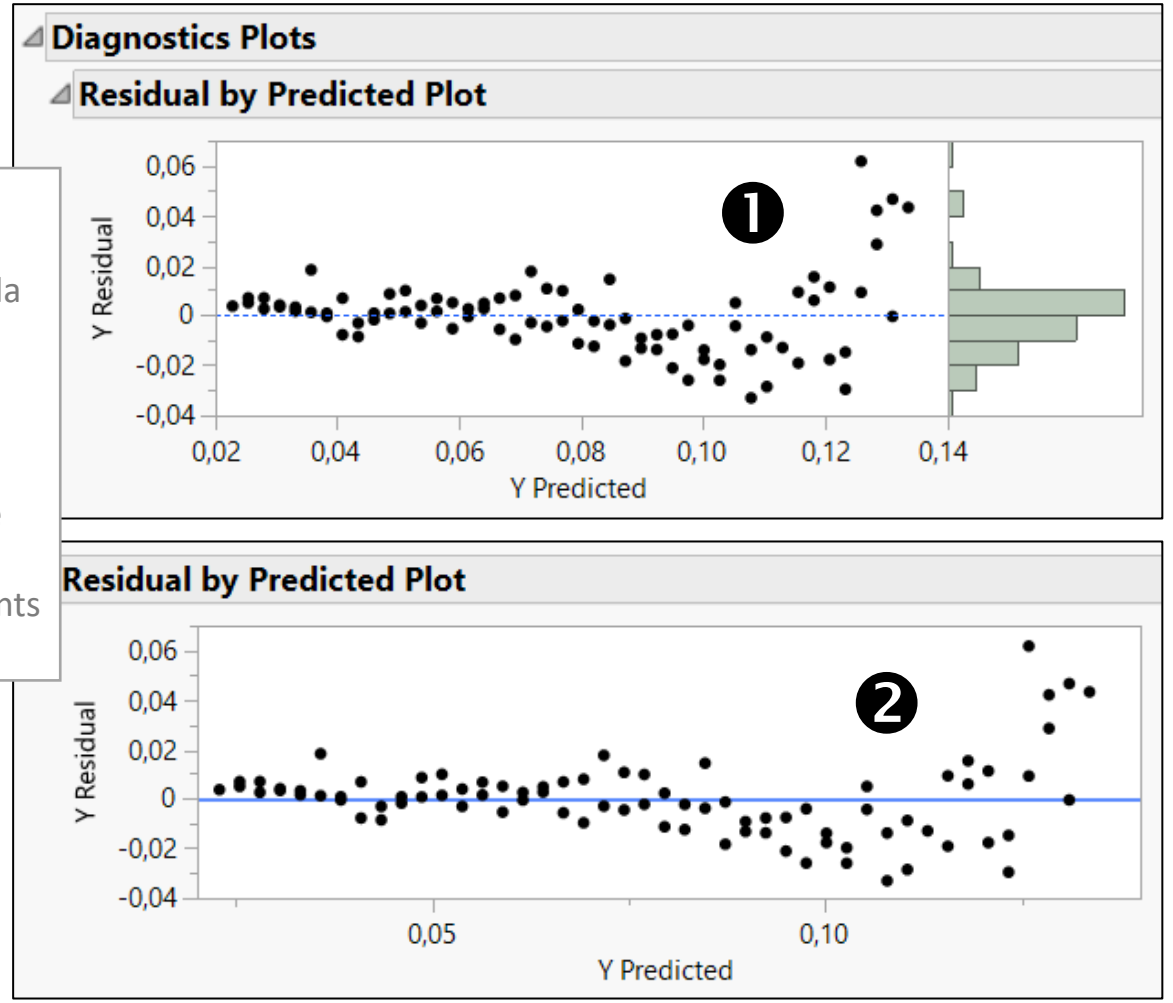
Les conseils de JMPy

Ici, la valeur des résidus semblent augmenter avec la valeur fittée. Phénomène connu sous le nom d'hétérosédasticité signifiant que la variabilité de la réponse change lorsque la valeur prédite augmente. Mince, il semblerait que notre régression ne soit pas parfaite!

Envie de rejoindre la communauté des heureux utilisateurs de JMP?

Un résidu, c'est la différence entre la valeur mesurée et la valeur prédite par ton modèle.

Vérifions ensemble l'**homoscédasticité** de nos résidus! Sont-ils indépendants de la valeur fittée?



Etape 3
Validation du modèle

3b. Résidus

Sont-ils:
- Homoscédastiques?
- Approx. norm. distribués?
- Indépendants?

L'analyse des résidus permet de s'assurer qu'il ne reste plus aucune information à extraire des résidus, que toutes les tendances sont bien incorporées dans le modèle et que seul le bruit de fond (aléatoire) demeure.

Plateformes JMP

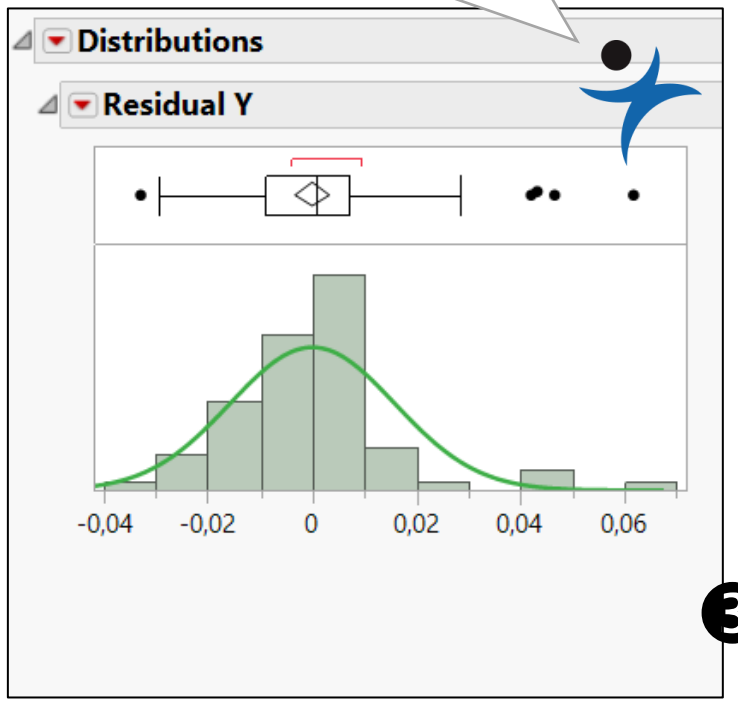
1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Régression simple (Linear Fit), puis sélectionne Tracer les résidus (Plot Residuals)
2. Via la plateforme Modèle Linéaire (Fit Model). ▼ Réponse (Response), puis Diagnostics de lignes (Row Diagnostics) > Tracer les résidus en fonction des quantiles normaux (Residual by Predicted Plot)
3. Sauvegarder les résidus dans une colonne et, via la plateforme Distributions, effectue un test de Shapiro-Wilk ou d'Anderson-Darling. ▼ Ajustement Continu (Continuous Fit) > Ajustement Normal puis ▼ Distribution normale ajustée (Fitted Normal Distribution) >Qualité de l'ajustement (Goodness of Fit)

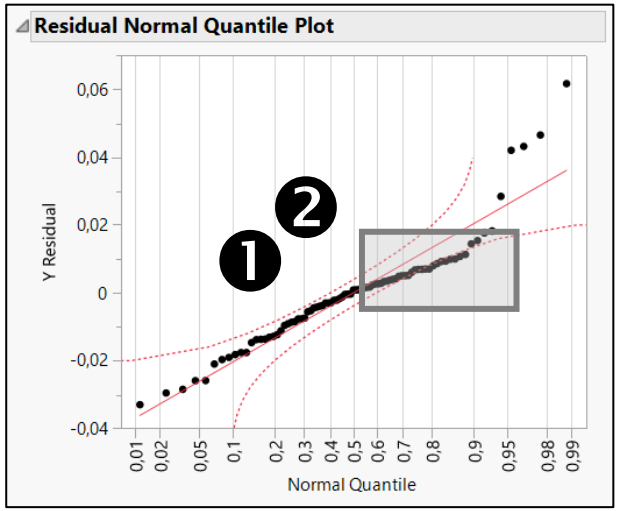
Les conseils de JMPy

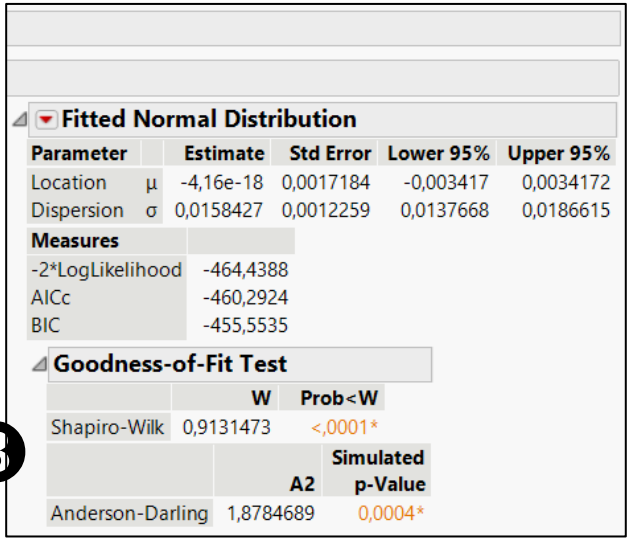
Le chart des quantiles normaux des résidus indique une légère sortie des limites. La faible p-value conduit à rejeter l'hypothèse d'une distribution normale.

Envie de rejoindre la communauté des heureux utilisateurs de JMP?

Vérifions ensemble que nos résidus sont **appromativement normalement distribués**. A moins d'une très grande déviation par rapport à la normalité ou d'un pattern particulier, il n'y a généralement pas à s'inquiéter outre mesure de cette étape.







Etape 3
Validation du modèle

3b. Résidus

Sont-ils:
- Homoscédastiques?
- Approx. norm. distribués?
- Indépendants?

L'analyse des résidus permet de s'assurer qu'il ne reste plus aucune information à extraire des résidus, que toutes les tendances sont bien incorporées dans le modèle et que seul le bruit de fond (aléatoire) demeure.

Plateformes JMP

1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Régression simple (Linear Fit), puis sélectionne Tracer les résidus (Residual by Row Plot)
2. Via la plateforme Modèle Linéaire (Fit Model). ▼ Réponse (Response), puis Diagnostics de lignes (Row Diagnostics) > Tracer les résidus en fonction des lignes (Residual by Row Plot)

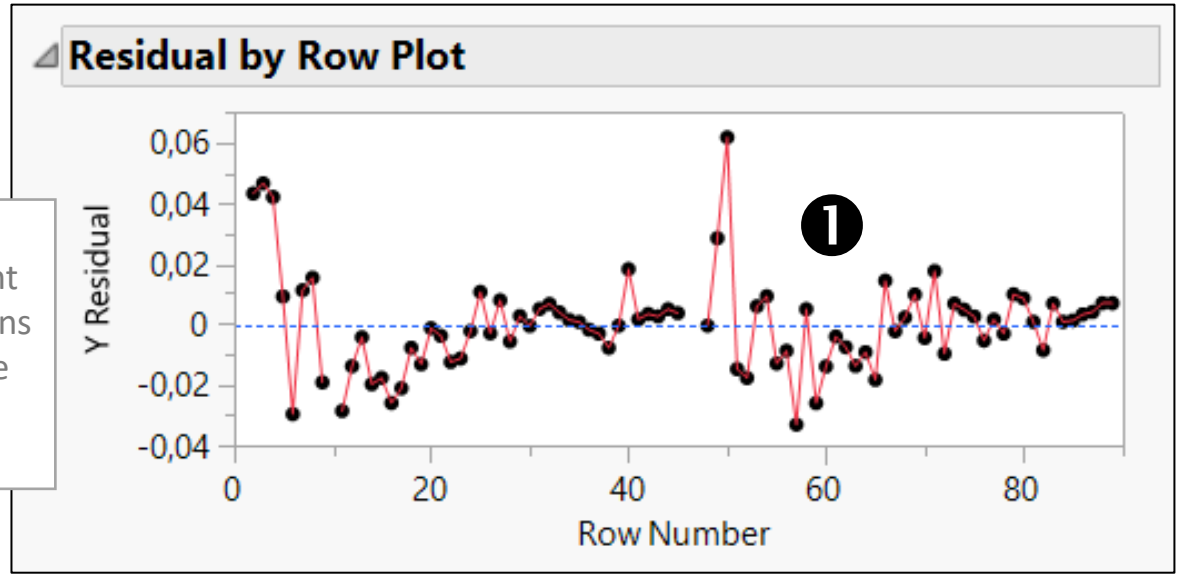
Les conseils de JMPy

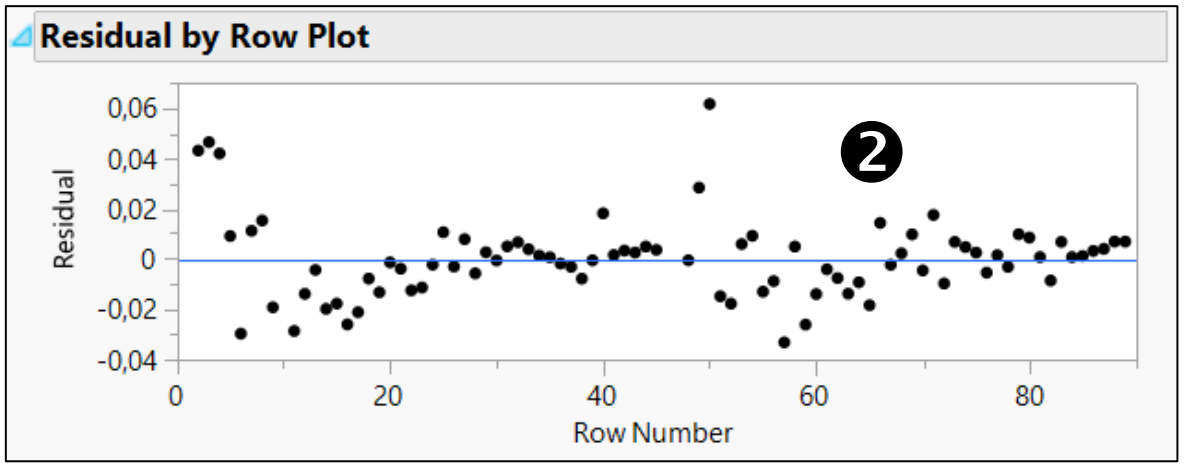
Observe le graphique des résidus en fonction de l'ordre des observations:
- Y-a-t-il une tendance?
- Sont-ils également répartis?
- Y-a-t-il autant d'observations au-dessus que en-dessous de 0.

Ici, le graphique ne semble indiquer rien ici d'anormal (pas de pattern)

Envie de rejoindre la communauté des heureux utilisateurs de JMP?

Vérifions ensemble que nos résidus sont **indépendants** les uns des autres (on parle aussi d'auto-corrélation)





Etape 3
Validation du modèle

3c. Outliers

Y-a-t-il des valeurs qui peuvent être considérées comme réellement aberrantes?

Parfois, il arrive qu'un seul individu atipyque fausse tous les résultats. Ils convient donc d'identifier ces potentielle valeurs aberrantes au sein du jeu de données.

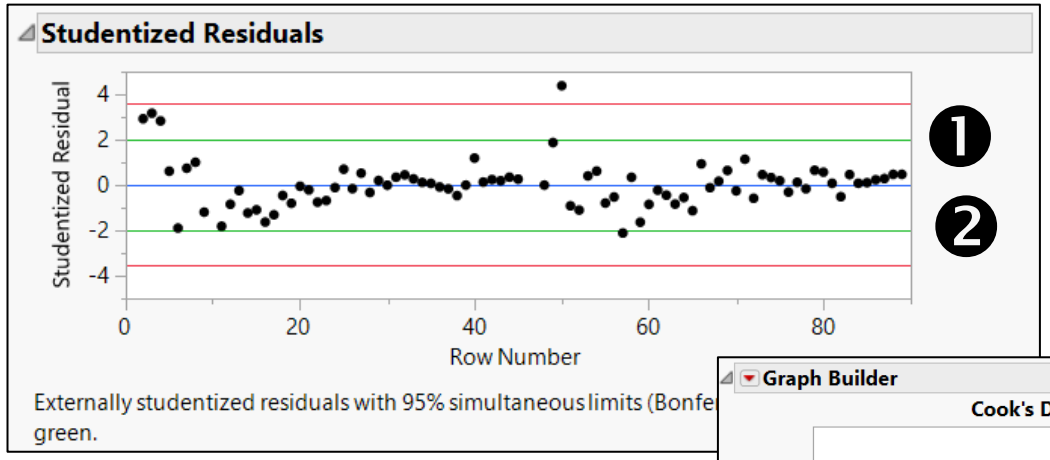
Plateformes JMP

1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Régression simple (Linear Fit), puis sélectionne Sauver les résidus Studentisés (Save Studentized Residuals)
2. Via la plateforme Modèle Linéaire (Fit Model). ▼ Réponse (Response), puis Diagnostics de lignes (Row Diagnostics) > > Tracer les résidus Studentisés (Plot Studentized Residuals)
3. Via la plateforme Modèle Linéaire (Fit Model). ▼ Réponse (Response), puis Diagnostics de lignes (Row Diagnostics) > Save Columns > Influence de Cook (Cook's D Influence).

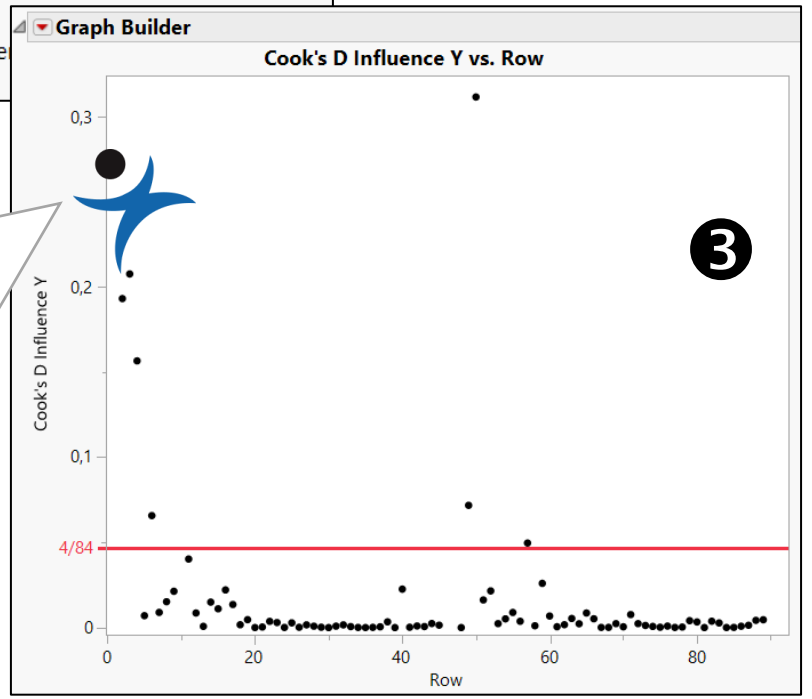
Les conseils de JMPy

Dans le graphique des résidus Studentisés, étudie les points expérimentaux supérieurs à |2.5| écart-types, dans celui des Distances de Cook ceux supérieurs à y=4/N. Attention, ne les écarte pas trop vite, cela peut être le signe de non linéarités. Dans notre cas, nous avons déjà des signes comme quoi notre modèle pourrait ne pas être linéaire. Passons plutôt à l'étape suivante.

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Les distances de Cook mesurent de combien les estimations du modèle changeraient si une observation était effacée du jeu de données. Tracer cette colonne en fonction de l'ordre des observations et ajouter la droite correspondante à y=4/N (N=taille de l'échantillon). Les valeurs au-dessus de cette droite sont normalement considérés comme fortement influente.



Etape 4
Transformation des données

4a. Non-Linéaire

Un modèle non-linéaire serait-il plus adapté? Essaie d'introduire les termes en X^2, X^3, etc.

Tu as des suspicions comme quoi ton modèle pourrait ne pas être linéaire? C'est bien possible! Commençons par introduire les non-linéarités au moyen de termes en X d'ordre supérieur.

Plateformes JMP

1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Ajustement polynomial (Fit Polynomial), puis sélectionne 2, 3, etc.
2. Via la plateforme Modèle Linéaire (Fit Model). Va jusqu'à l'ordre 5 (par exemple) et observe les résultats.

Les conseils de JMPy

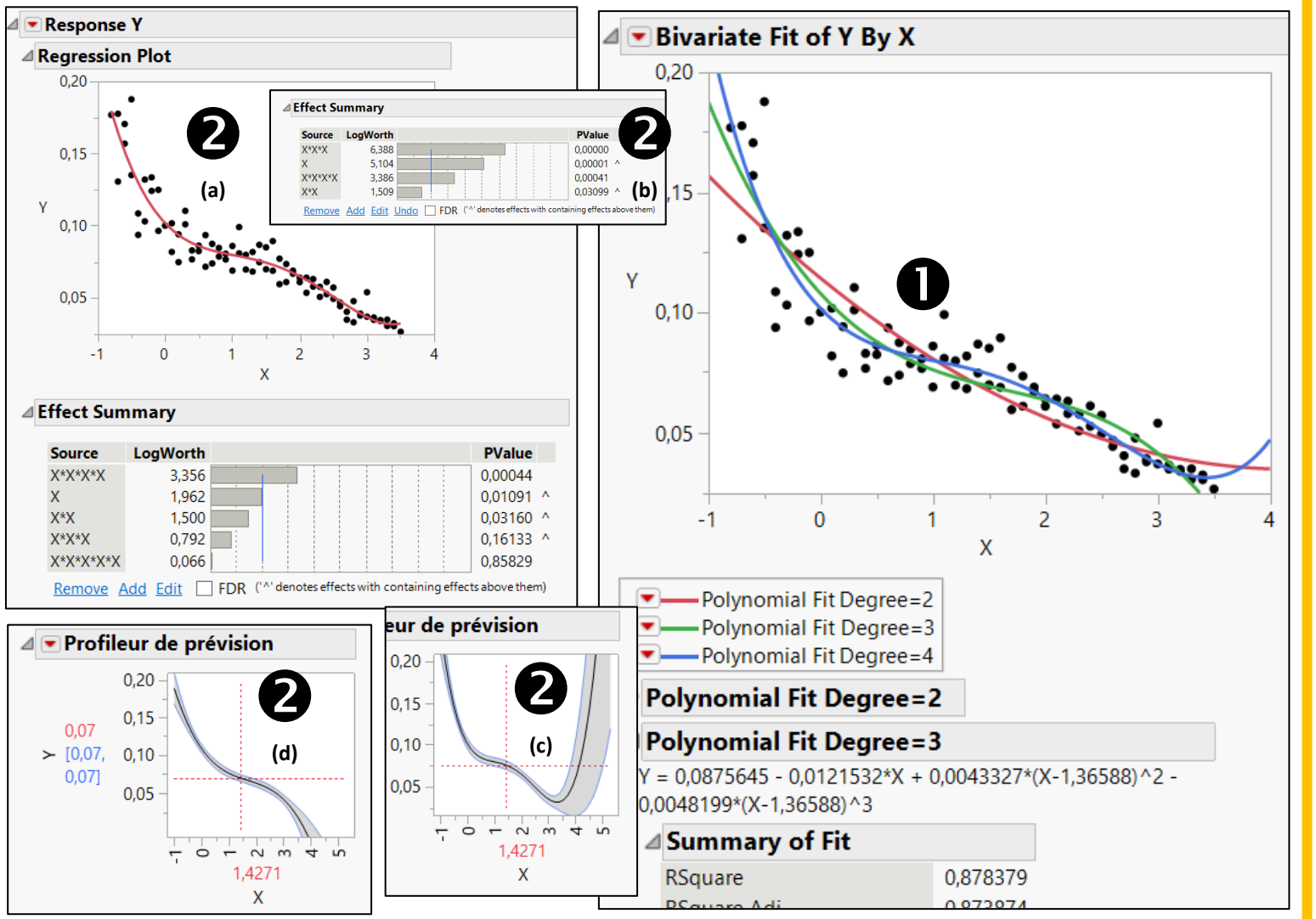
Dans la première plateforme, un modèle non linéaire du 3ème ordre semble visuellement mieux convenir. Dans la seconde, on voit que:

- La p-value de l'ordre 5 indique que ce terme n'est pas significatif, il peut être retirer (2a)
- Après avoir retirer ce terme (2b), la p-value de l'ordre 4 est significative. Le profileur montrant un comportement visiblement non adapté (2c), on se limitera à l'ordre 3 (2d). (On se serait posé la question à l'étape 5b sinon).

Avec cet ajout, le pouvoir explicatif R^2 passe de 80.4% à 87.9%

N'oublie pas de valider le modèle en retournant à l'étape 3!

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 4
Transformation des données

4b. X-Transf.

Une transformation de X permet-elle une meilleure modélisation? Essaie 1/X, Log(X), X^0.5, etc.

Parfois, introduire les non-linéarités au moyen de termes en X d'ordre supérieur n'est pas suffisant. Il peut être intéressant de tester des transformations de la variable X au moyen de fonctions divers: Log, Sqrt, Réciproque, etc.

Plateformes JMP

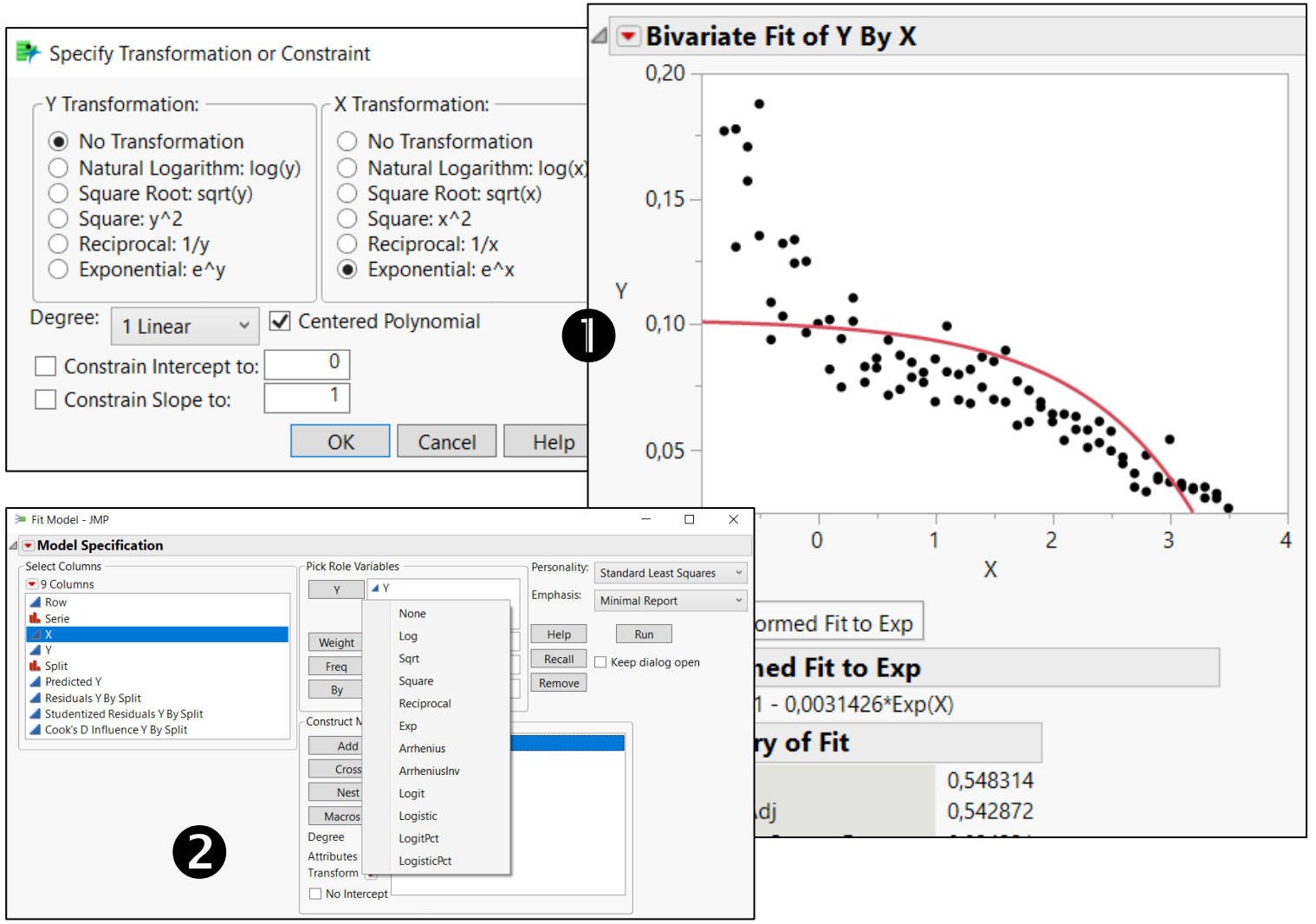
1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Autre ajustement (Fit Special)
2. Via la plateforme Modèle Linéaire (Fit Model). Sélectionne X, cliquer sur ▼ Transformer et sélectionner une fonction de transformation.

Les conseils de JMPy

Sélectionne une transformation applicable (si X<0, Log, Sqrt, etc sont exclues) ainsi qu'un degré (démarre avec 1!!). Dans le cas présent, aucune transformation n'améliore la modélisation.

N'oublie pas de valider le modèle en retournant à l'étape 3!

Envie de rejoindre la communauté des heureux utilisateurs de JMP?



Etape 4
Transformation des données

4b. Y-Transf.

Une transformation de Y permet-elle une meilleure modélisation? Essaie 1/Y, Log(Y), Y^0.5, etc.

Parfois, introduire les non-linéarités uniquement sur X n'est pas suffisant. Il peut être intéressant de tester conjointement des transformations de la variable Y au moyen de fonctions divers: Log, Sqrt, Réciproque, etc.

Plateformes JMP

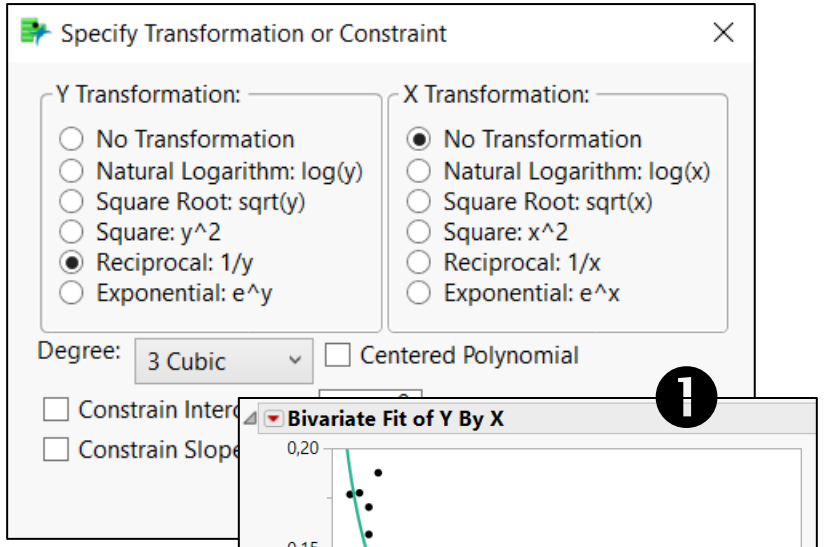
1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y), ▼ Autre ajustement (Fit Special)
2. Via la plateforme Modèle Linéaire (Fit Model). Sélectionne X, cliquer sur ▼ Transformer et sélectionner une fonction de transformation.

Les conseils de JMPy

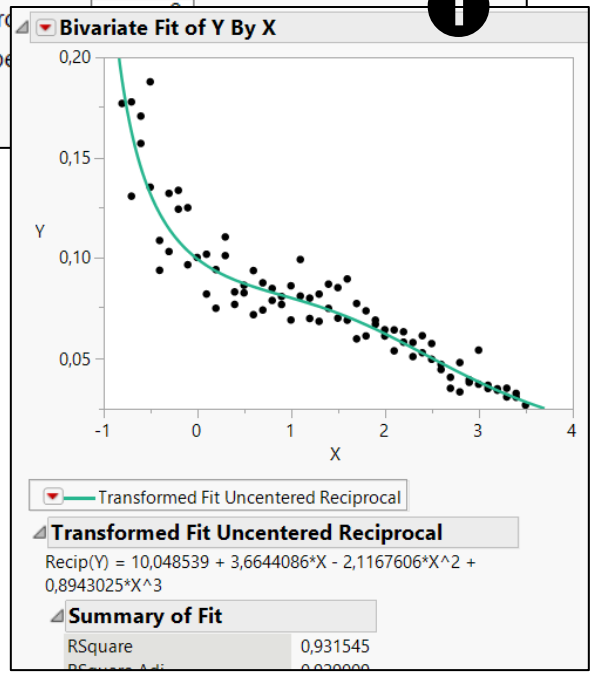
Pour la première plateforme, sélectionne une transformation applicable pour Y, pour X ainsi qu'un degré. Pour notre cas, une transformation de Y en 1/Y et un polynôme du 3ème ordre en X conduit à un pouvoir explicatif de 93.1%. Pour la seconde plateforme, la transformation de Y en 1/Y conduit à la même modélisation, avec de plus la significativité des termes qui permet de conclure rapidement qu'il convient d'aller jusqu'à l'ordre 3.

N'oublie pas de valider le modèle en retournant à l'étape 3!

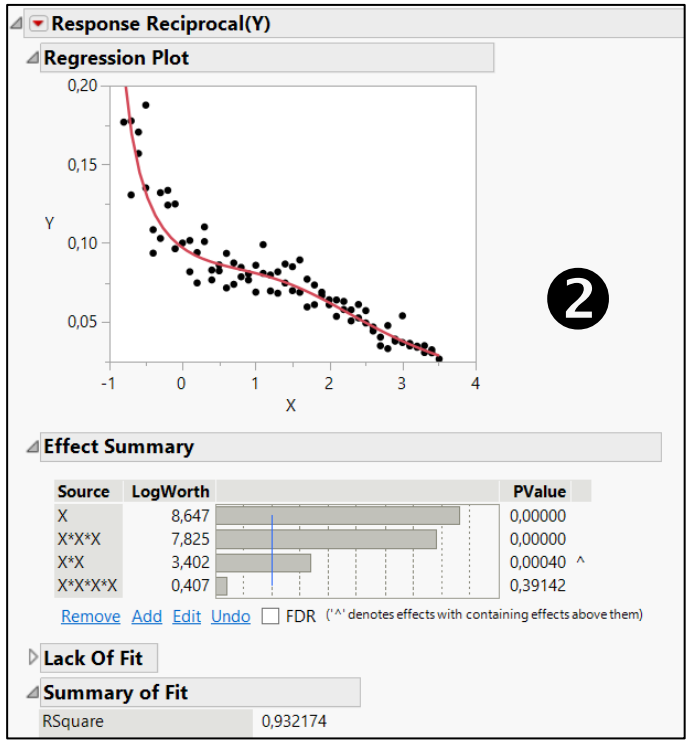
Envie de rejoindre la communauté des heureux utilisateurs de JMP?



❶



❷



En fonction des données, la transformation de Box-Cox peut donner une idée de la transformation de Y à appliquer.

Etape 5
Ultimes étapes de validation

5a. Suspects?

Les points écartés à l'étape 1b sont-ils vraiment suspects? Que se passe-t-il si tu les rajoutes?

Tu te rappelles des points que tu avais écarté de l'analyse à l'étape 1b? Et si on les rajoutait maintenant, juste pour voir ce que cela donne. Sont-ils toujours si suspects?

Plateformes JMP

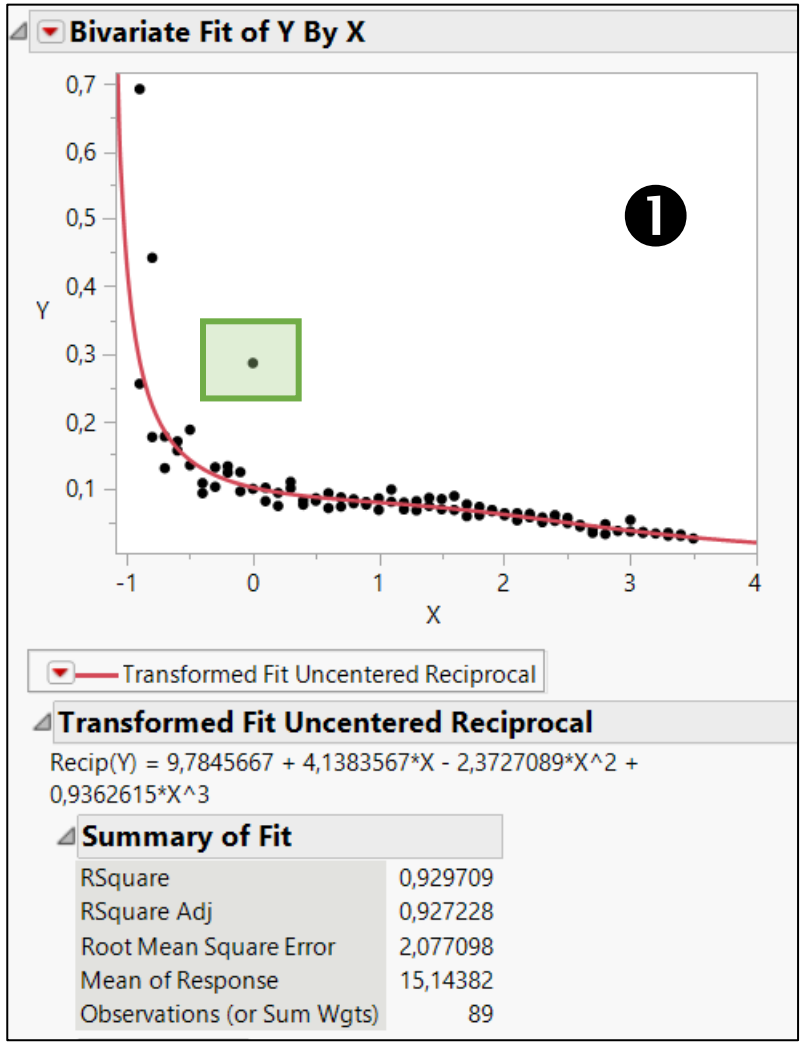
1. Via la plateforme Ajuster Y en fonction de X (Fit X by Y)
2. Via la plateforme Modèle Linéaire (Fit Model)

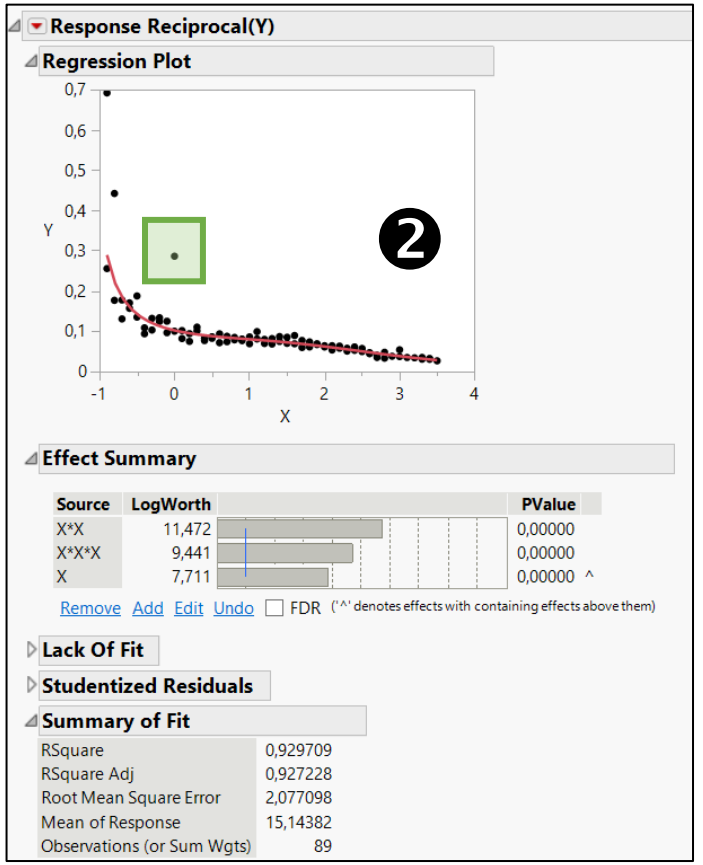
Les conseils de JMPy

Dans le cas présent, une transformation de Y en 1/Y et un polynôme du 3ème ordre en X conduit à un pouvoir explicatif de 92.9% sur l'ensemble des points du jeu de données. Suis de nouveau les étapes de validation du modèle … finalement, il n'y avait peut-être qu'une seule valeur aberrante!

N'oublie pas de valider le modèle en retournant à l'étape 3!

Envie de rejoindre la communauté des heureux utilisateurs de JMP?





Etape 5
Ultimes étapes de validation

5b. Sens?

Finalement, ton modèle a-t-il un sens? Es-tu capable d'expliquer son comportement?

Finalement, après tout ce travail, il reste une ultime question à se poser! Mais c'est probablement la plus importante de toute! Le modèle que tu as créé fait-il sens?

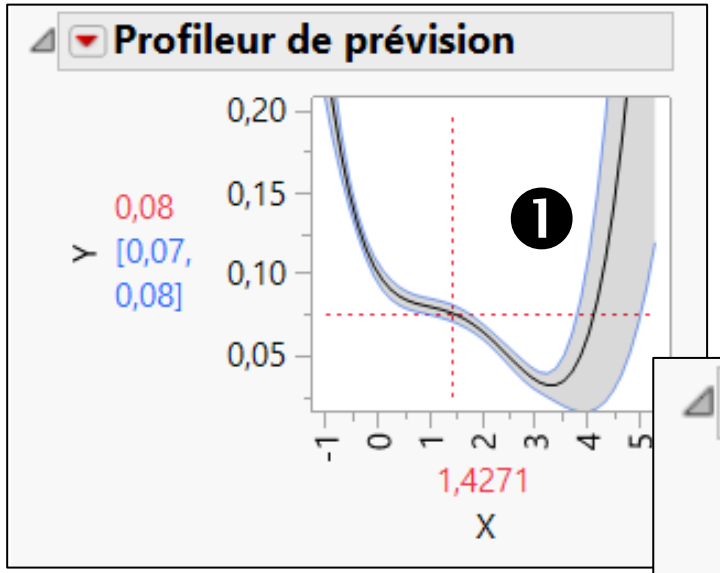
Plateformes JMP

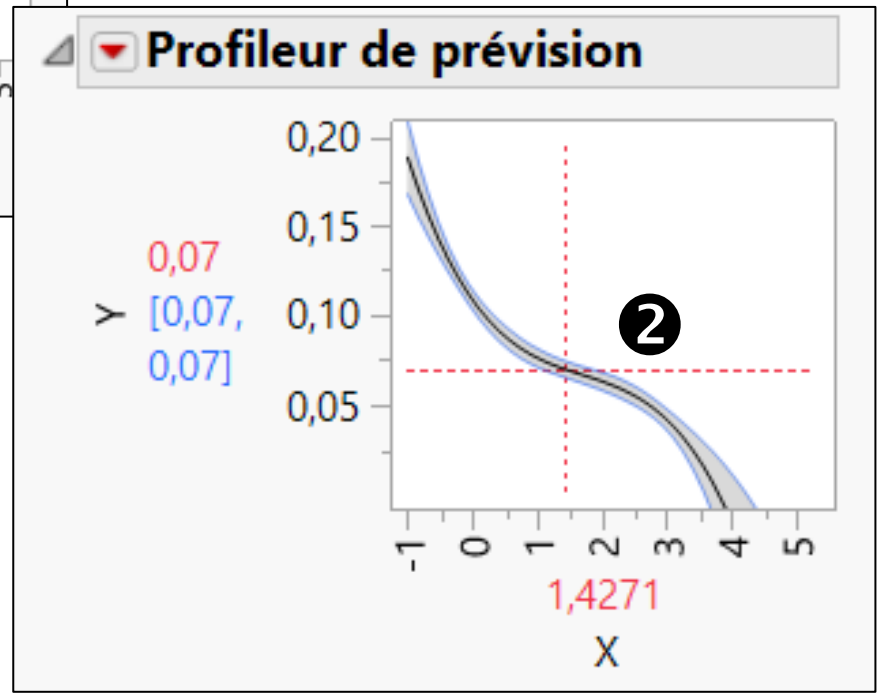
Aucune

Les conseils de JMPy

Nous arrivons au terme du workflow, il reste à se poser l'ultime question. Ne la néglige surtout pas, c'est la plus importante de toutes les questions que je t'ai posée … ton modèle a-t-il un sens? Es-tu capable d'expliquer son comportement? Pour illustrer cette question, imagine que tu n'aies pas retiré l'ordre 4 en X précédemment. Tu n'aurais alors pas le modèle avec un comportement comme décrit par le profiler (2) mais comme (1). Il te faudrait alors t'interroger: « Quand X>=3.5, physiquement, est-il logique que Y augmente de la sorte? Ne doit-il pas continuer de diminuer? »

Envie de rejoindre la communauté des heureux utilisateurs de JMP?





Etape 5
Ultimes étapes de validation

5c. Happy!

Félicitations! Tu es arrivé(e) au bout du workflow!

Toutes mes félicitations! Tu es arrivé(e) au bout du workflow! Tu devrais être rassuré sur ton modèle désormais.

Plateformes JMP

Aucune

Les conseils de JMPy

Bravo pour ton parcours! N'hésite pas à le refaire pour bien comprendre.

Pour l'exemple que nous avons étudié, je te donne la réponse théorique ci-contre. En suivant pas à pas le workflow, nous sommes quasiment arrivés à la retrouver.

Envie de rejoindre la communauté des heureux utilisateurs de JMP?

Pour notre cas, je te rappelle le modèle que nous avons trouvé ensemble

$$Recip(Y) = 9{,}7845667 + 4{,}1383567*X - 2{,}3727089*X^2 + 0{,}9362615*X^3$$

… et voici ce qu'il fallait trouver. Pas mal du tout!

$$\left(\frac{1}{\left(\left(10+5\bullet X+-3\bullet X^{2}+1\bullet X^{3}\right)+\text{Random Normal}\left(0,1\right)\right)}\right)\bullet\left(\begin{matrix}\text{Row}\left(\right)==10\Rightarrow 3\\ \text{else}\qquad\Rightarrow 1\end{matrix}\right)\bullet\text{Random Normal}\left(1,0{,}1\right)$$

… ça, c'était notre point aberrant!

Merci de nous avoir suivi jusqu'ici

Vous avez aimé cette étude de cas?
Vous avez aimé JMPy?
Vous souhaitez qu'il revienne?

Dites-le nous sur le site de
La communauté des utilisateurs francophones de JMP ➲