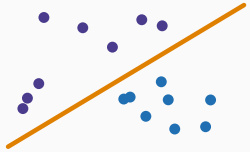
Analyse quantitative

Chapitre 9 : Les diagnostics de régression



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 29 août 2026

Plan du chapitre

Avant de commencer

L'hétéroscédasticité

La multicolinéarité

Biais-variance, le choix

Synthèse

Avant de commencer

Pourquoi ce chapitre ?

Les chapitres 7 et 8 ont supposé les hypothèses des MCO satisfaites. Ce chapitre examine ce qui se passe lorsqu'elles ne le sont pas : **hétéroscédasticité**, **multicolinéarité**, mauvaise spécification. On y voit comment détecter ces problèmes, comment y remédier, et comment choisir entre plusieurs modèles concurrents — question qui débouche sur l'arbitrage **biais-variance**.

L’hétéroscédasticité

Nature du problème

Définition

Il y a **hétéroscédasticité** lorsque la variance des erreurs n'est pas constante : $V[\varepsilon_i \mid X_i]$ dépend de X_i. L'hypothèse d'homoscédasticité des MCO est alors violée.

Les conséquences — et ce qui reste valide

Les estimateurs MCO restent **sans biais** et convergents : les coefficients ne sont pas faussés. En revanche, leurs **erreurs types** sont mal estimées, ce qui invalide les tests t et F. Les MCO perdent aussi leur optimalité : ils ne sont plus BLUE. Le problème porte donc sur l'**inférence**, non sur l'estimation ponctuelle.

Un phénomène systématique en finance

L'hétéroscédasticité est la règle sur données financières : la volatilité des rendements varie fortement dans le temps, alternant périodes calmes et périodes agitées. C'est ce qu'on appelle l'**agrégation de volatilité**, et c'est ce que les modèles GARCH cherchent à modéliser explicitement.

Détection et remèdes

Tester l'hétéroscédasticité

Les tests de **White** et de **Breusch-Pagan** régressent les résidus au carré sur les variables explicatives (et, pour White, sur leurs carrés et produits croisés). Si ces variables expliquent significativement la dispersion des résidus, l'hypothèse d'homoscédasticité est rejetée. Un examen graphique des résidus contre les valeurs ajustées reste un premier diagnostic très efficace.

Trois remèdes

Les **erreurs types robustes** (dites de White) corrigent l'inférence sans modifier les coefficients : c'est la solution la plus courante, et beaucoup de praticiens les utilisent systématiquement. Les **moindres carrés pondérés** repondèrent les observations par l'inverse de leur variance, plus efficaces mais exigeant de connaître la forme de l'hétéroscédasticité. Enfin, une **transformation** de la variable expliquée — souvent le logarithme — stabilise parfois la variance.

La multicolinéarité

Distinguer deux situations

Colinéarité parfaite et multicolinéarité

La **colinéarité parfaite** — une variable exactement combinaison linéaire des autres — rend l'estimation **impossible** : les coefficients ne sont pas identifiables. La **multicolinéarité** désigne une corrélation forte mais imparfaite : l'estimation reste possible, mais devient imprécise.

Les symptômes

Coefficients de **grande variance**, donc erreurs types élevées et t faibles ; forte **instabilité** des estimations lorsqu'on ajoute ou retire une observation ; signes parfois contraires à l'intuition. Signal caractéristique : un R^2 élevé et un test de Fisher très significatif alors qu'**aucun** coefficient n'est individuellement significatif.

Mesurer et traiter

Le **facteur d'inflation de la variance** (VIF) quantifie le phénomène ; une valeur supérieure à 10 est généralement jugée problématique. Les remèdes : retirer une des variables redondantes, les combiner en un indice, recourir à l'**analyse en composantes principales** (chapitre 14), ou augmenter l'échantillon. Mais si l'objectif est la **prévision** et non l'interprétation des coefficients, la multicolinéarité n'est pas gênante.

Biais-variance, le choix

Biais et variance

Un modèle **trop simple** omet des variables pertinentes : il est **biaisé** mais stable. Un modèle **trop riche** capte le bruit de l’échantillon : il est peu biaisé mais de forte **variance**, donc instable d’un échantillon à l’autre. L’erreur de prévision totale combine ces deux sources, et la minimiser exige de trouver le point d’équilibre.

Sous-ajustement et sur-ajustement

Le **sous-ajustement** manque une structure réelle ; le **sur-ajustement** modélise du bruit comme s’il était du signal. Ce dernier se reconnaît à un contraste marqué entre une performance excellente dans l’échantillon et médiocre hors échantillon. C’est le fil conducteur des chapitres 14 et 15.

Deux procédures de sélection

Sélection descendante et ascendante

La procédure **descendante** (*general to specific*) part du modèle le plus large et retire successivement la variable la moins significative jusqu'à ce que toutes le soient. La procédure **ascendante** (*specific to general*) part du modèle minimal et ajoute la variable la plus significative tant que l'apport le justifie.

Les critères d'information

Les critères **AIC** et **BIC** arbitrent formellement entre ajustement et complexité, en pénalisant le nombre de paramètres. Le **BIC** pénalise plus lourdement que l'AIC et sélectionne donc des modèles plus parcimonieux — préférable lorsque l'objectif est l'interprétation.

Une mise en garde

Ces procédures enchaînent de nombreux tests sur les mêmes données : elles retombent exactement sur le problème des **tests multiples** du chapitre 6. Les niveaux de significativité affichés à l'issue d'une sélection automatique sont donc trompeurs, et le modèle retenu doit impérativement être validé **hors échantillon**.

Synthèse

Ce qu'il faut retenir

Synthèse du chapitre

L'**hétéroscédasticité** ne biaise pas les coefficients mais fausse leurs erreurs types, invalidant les tests ; elle est la règle sur données financières, et les **erreurs types robustes** en sont le remède standard. La **colinéarité parfaite** empêche toute estimation, tandis que la **multicolinéarité** la rend imprécise — reconnaissable à un R^2 élevé sans coefficient individuellement significatif, mesurable par le **VIF**, et sans conséquence si l'objectif est la seule prévision. Le choix de modèle relève de l'arbitrage **biais-variance** : trop simple, le modèle est biaisé ; trop riche, il capte le bruit. Les procédures de sélection ascendante ou descendante et les critères **AIC/BIC** organisent ce compromis, mais enchaînent des tests multiples : seule la validation **hors échantillon** confirme réellement un modèle.