Test Régression Linéaire - Niveau 2 (Avancé)

UniBank Haiti - Data Analyst

Durée: 45 minutes
Questions: 25
Niveau: Avancé
Sujets: Diagnostics avancés, multicolinéarité, régression logistique, interprétation

---

Q1. Dans un modèle de scoring crédit avec 8 variables, le $R^2 = 0.75$ et le R^2 ajusté = 0.68. Que suggère cette différence?

A) Le modèle est excellent

B) Il y a probablement des variables inutiles dans le modèle

C) Le modèle souffre de sous-ajustement

D) Les données sont parfaitement normales

Réponse: B) Il y a probablement des variables inutiles dans le modèle

Un écart important entre R^2 et R^2 ajusté (7 points ici) indique que certaines variables n'apportent pas de pouvoir explicatif réel. Le R^2 ajusté pénalise l'ajout de variables inutiles.

Q2. Vous observez un VIF de 12.5 pour la variable "revenus_mensuels" dans votre modèle. Quelle est l'action appropriée?

A) Ignorer car le VIF est toujours élevé pour les revenus

B) Investiguer les corrélations avec d'autres variables et potentiellement en supprimer

C) Multiplier les revenus par 12.5 pour corriger

D) Ajouter plus de variables au modèle

Réponse: B) Investiguer les corrélations avec d'autres variables et potentiellement en supprimer

VIF > 10 indique une multicolinéarité sévère. Il faut identifier quelle(s) variable(s) est(sont) fortement corrélée(s) avec les revenus (ex: patrimoine, type de logement) et décider laquelle supprimer ou combiner.

Q3. Le test de Durbin-Watson donne une statistique DW = 0.8 pour un modèle de prévision des défauts de paiement. Comment interprétez-vous ce résultat?

A) Pas d'autocorrélation, le modèle est valide

B) Autocorrélation positive des résidus - hypothèse d'indépendance violée

C) Autocorrélation négative des résidus

D) Hétéroscédasticité confirmée

Réponse: B) Autocorrélation positive des résidus - hypothèse d'indépendance violée

DW ≈ 2 = pas d'autocorrélation. DW < 2 (ici 0.8) indique une autocorrélation positive. Cela suggère que les observations ne sont pas indépendantes, souvent problématique pour des données temporelles.

Q4. Dans une régression logistique pour le scoring, le coefficient β de "ancienete_emploi" est 0.15 (p < 0.01). Quel est l'odds ratio et son interprétation?

A) OR = 0.15, chaque année augmente les chances d'approbation de 15%

B) OR = 1.16, chaque année multiplie les chances d'approbation par 1.16

C) OR = 0.85, chaque année diminue les chances d'approbation

D) L'odds ratio ne peut pas être calculé sans l'intercept

Réponse: B) OR = 1.16, chaque année multiplie les chances d'approbation par 1.16

OR = e^β = e^0.15 ≈ 1.16. L'interprétation: pour chaque année supplémentaire d'ancienneté, les odds (chances relatives) d'approbation sont multipliés par 1.16, soit une augmentation de 16%.

Q5. Vous effectuez un test de Breusch-Pagan et obtenez p = 0.002. Quelle hypothèse LINE est violée?

A) Linéarité

B) Indépendance

C) Normalité

D) Égalité des variances (homoscédasticité)

Réponse: D) Égalité des variances (homoscédasticité)

Le test de Breusch-Pagan teste l'hétéroscédasticité. p < 0.05 rejette H_0 (homoscédasticité), donc la variance des résidus n'est pas constante - violation de l'hypothèse E(galité des variances).

Q6. Dans un modèle de prédiction du montant de prêt accordé, vous avez: Montant = 50000 + 0.3×Revenu - 5000×Nb_Credits_Actifs. Pour un client avec revenu de 200,000 HTG et 2 crédits actifs, quel montant prédit?

A) 100,000 HTG

B) 110,000 HTG

C) 100,500 HTG

D) 50,000 HTG

Réponse: A) 100,000 HTG

Montant = 50000 + 0.3(200000) - 5000(2) = 50000 + 60000 - 10000 = 100,000 HTG

Q7. Quelle transformation est appropriée si vos résidus montrent une distribution fortement asymétrique à droite?

A) Transformation logarithmique de Y

B) Standardisation de X

C) Ajout de variables polynomiales

D) Suppression des outliers uniquement

Réponse: A) Transformation logarithmique de Y

Une asymétrie positive des résidus suggère souvent que la variable dépendante devrait être transformée en log(Y). Ceci "compresse" les grandes valeurs et peut normaliser les résidus.

Q8. Le coefficient de la variable "score_interne" dans votre modèle a un intervalle de confiance à 95% de [-0.02, 0.15]. Cette variable est-elle significative au seuil α = 0.05?

A) Oui, car l'intervalle est étroit

B) Non, car l'intervalle contient 0

C) Oui, car la borne supérieure est positive

D) Impossible à déterminer sans la p-value

Réponse: B) Non, car l'intervalle contient 0

Si l'IC à 95% contient 0, cela signifie qu'on ne peut pas rejeter H_0: $\beta=0$ au seuil 5%. Le coefficient n'est pas significativement différent de 0.

Q9. Dans une régression Ridge (L2), que se passe-t-il quand le paramètre λ augmente?

A) Les coefficients augmentent vers l'infini

B) Les coefficients se rapprochent de 0 mais ne deviennent jamais exactement 0

C) Tous les coefficients deviennent exactement 0

D) Le R^2 augmente systématiquement

Réponse: B) Les coefficients se rapprochent de 0 mais ne deviennent jamais exactement 0

Ridge (L2) shrink les coefficients vers 0 mais ne les annule jamais complètement. C'est la différence avec Lasso (L1) qui peut mettre des coefficients exactement à 0.

Q10. Vous comparez deux modèles de scoring: Modèle A (AIC=1250, BIC=1280) et Modèle B (AIC=1220, BIC=1240). Lequel préférer?

A) Modèle A car AIC et BIC sont plus élevés

B) Modèle B car AIC et BIC sont plus faibles

C) On ne peut pas décider sans le R^2

D) On préfère toujours le modèle avec plus de variables

Réponse: B) Modèle B car AIC et BIC sont plus faibles

Pour AIC et BIC, plus faible = meilleur. Ces critères pénalisent la complexité du modèle. Le Modèle B ayant les deux critères plus bas est préférable.

Q11. Dans votre modèle, la statistique F globale a une p-value de 0.35. Que concluez-vous?

A) Au moins une variable est significative

B) Aucune variable n'est significative - le modèle n'explique rien

C) Le modèle est parfait

D) Il faut ajouter plus de variables

Réponse: B) Aucune variable n'est significative - le modèle n'explique rien

Le test F global teste si au moins un coefficient est différent de 0. Avec $p = 0.35 > 0.05$, on ne rejette pas H_0, donc collectivement les variables n'expliquent pas significativement Y.

Q12. Un graphique des résidus vs valeurs prédites montre une forme en "entonnoir" (plus de dispersion à droite). Quel problème cela indique-t-il?

A) Non-linéarité

B) Hétéroscédasticité

C) Multicolinéarité

D) Autocorrélation

Réponse: B) Hétéroscédasticité

La forme en entonnoir (fan shape) où la variance augmente avec les valeurs prédites est le signe classique d'hétéroscédasticité - la variance n'est pas constante.

Q13. Pour la régression logistique, quelle fonction lie la probabilité aux variables explicatives?

A) Fonction linéaire

B) Fonction logit (log-odds)

C) Fonction exponentielle

D) Fonction quadratique

Réponse: B) Fonction logit (log-odds)

logit(p) = log(p/(1-p)) = $\beta_0 + \beta_1 X_1 + \ldots$ La régression logistique modélise le log des odds comme fonction linéaire des prédicteurs.

Q14. Dans un modèle de prédiction de LGD (Loss Given Default), le coefficient de "presence_garantie" est -0.25. Comment l'interpréter?

A) La présence de garantie augmente la LGD de 25%

B) La présence de garantie diminue la LGD de 0.25 unité (ou 25 points de pourcentage si LGD en %)

C) La garantie n'a pas d'effet

D) Il faut transformer le coefficient en odds ratio

Réponse: B) La présence de garantie diminue la LGD de 0.25 unité (ou 25 points de pourcentage si LGD en %)

Dans une régression linéaire simple, β = -0.25 signifie que passer de 0 (pas de garantie) à 1 (garantie) diminue Y de 0.25. Si LGD est en pourcentage, cela représente -25pp.

Q15. Quelle méthode utilisez-vous pour sélectionner les variables dans un modèle avec 50 candidats?

A) Inclure toutes les 50 variables

B) Utiliser uniquement l'intuition métier

C) Stepwise selection (forward, backward, ou bidirectionnel) + validation croisée

D) Choisir les 5 premières variables alphabétiquement

Réponse: C) Stepwise selection (forward, backward, ou bidirectionnel) + validation croisée

Avec beaucoup de variables candidates, une sélection systématique (stepwise) guidée par AIC/BIC, combinée à la validation croisée pour éviter l'overfitting, est l'approche standard.

Q16. Le pseudo-R^2 de McFadden de votre régression logistique est 0.35. Est-ce acceptable?

A) Non, il devrait être > 0.9

B) Oui, 0.2-0.4 est généralement considéré comme un bon ajustement pour McFadden

C) Non, seul $R^2 > 0.5$ est acceptable

D) Le pseudo-R^2 n'a aucune signification

Réponse: B) Oui, 0.2-0.4 est généralement considéré comme un bon ajustement pour McFadden

Le pseudo-R^2 de McFadden ne s'interprète pas comme le R^2 classique. Des valeurs entre 0.2 et 0.4 sont généralement considérées comme représentant un bon ajustement.

Q17. Vous observez que la variable "code_postal" a une très forte importance dans votre modèle de scoring. Quelle préoccupation cela soulève-t-il?

A) Aucune, c'est une variable utile

B) Risque de proxy pour discrimination géographique/ethnique

C) Le code postal devrait être normalisé

D) Il faut ajouter plus de codes postaux

Réponse: B) Risque de proxy pour discrimination géographique/ethnique

Le code postal peut être corrélé avec l'origine ethnique ou le niveau socio-économique. Son utilisation importante dans un modèle de scoring peut introduire une discrimination indirecte (proxy discrimination).

Q18. Dans une analyse de sensibilité, vous doublez les valeurs de "montant_demande" et observez que les prédictions changent de façon non proportionnelle. Que suggère ce comportement?

A) Le modèle est parfaitement linéaire

B) Il y a des interactions ou des effets non-linéaires non capturés

C) Les données sont normalement distribuées

D) Le modèle a un R^2 de 1

Réponse: B) Il y a des interactions ou des effets non-linéaires non capturés

Dans un modèle linéaire pur, doubler X devrait approximativement doubler l'effet sur Y ($\times\beta$). Un comportement non proportionnel suggère des non-linéarités ou des interactions.

Q19. Quelle est la différence entre régression Ridge et Lasso pour la sélection de variables?

A) Aucune différence

B) Ridge met certains coefficients exactement à 0, Lasso non

C) Lasso met certains coefficients exactement à 0, Ridge non

D) Les deux mettent tous les coefficients à 0

Réponse: C) Lasso met certains coefficients exactement à 0, Ridge non

Lasso (L1) peut produire des coefficients exactement nuls, effectuant ainsi une sélection de variables automatique. Ridge (L2) shrink vers 0 mais ne les annule jamais complètement.

Q20. Le test de Shapiro-Wilk sur vos résidus donne p = 0.12. Que concluez-vous sur l'hypothèse de normalité?

A) Les résidus ne sont pas normaux

B) Les résidus sont parfaitement normaux

C) On ne rejette pas l'hypothèse de normalité (résidus approximativement normaux)

D) Le test est non concluant

Réponse: C) On ne rejette pas l'hypothèse de normalité (résidus approximativement normaux)

Avec p = 0.12 > 0.05, on ne rejette pas H_0 (normalité). Les résidus peuvent être considérés comme approximativement normaux, satisfaisant l'hypothèse N de LINE.

Q21. Dans un modèle de PD (Probability of Default), vous obtenez un coefficient de 0.5 pour "nb_retards_paiement". Quel est l'odds ratio?

A) 0.5

B) 1.5

C) 1.65 (e^0.5)

D) 2.0

Réponse: C) 1.65 (e^0.5)

OR = e^β = e^0.5 ≈ 1.649. Chaque retard supplémentaire multiplie les odds de défaut par environ 1.65, soit une augmentation de 65% des chances relatives de défaut.

Q22. Comment traiter la multicolinéarité parfaite entre “revenus_net” et “revenus_brut - impots”?

A) Garder les deux variables

B) Supprimer l’une des deux variables (celle moins pertinente métier)

C) Multiplier les deux par 0.5

D) Ajouter une troisième variable liée

Réponse: B) Supprimer l’une des deux variables (celle moins pertinente métier)

La multicolinéarité parfaite (une variable est combinaison linéaire exacte de l’autre) rend la matrice X’X non inversible. Il faut obligatoirement supprimer une des variables redondantes.

Q23. Vous construisez un modèle pour estimer le LTV (Lifetime Value) client. La variable “age_client” a un coefficient quadratique négatif (age^2). Que signifie cela?

A) L’effet de l’âge sur LTV est linéairement croissant

B) L’effet de l’âge sur LTV est en forme de U inversé (maximum à un certain âge)

C) L’âge n’a aucun effet

D) Les clients plus âgés ont toujours un LTV plus élevé

Réponse: B) L’effet de l’âge sur LTV est en forme de U inversé (maximum à un certain âge)

Un coefficient négatif pour le terme quadratique (age^2) avec un coefficient positif pour age crée une parabole inversée. Le LTV augmente avec l’âge jusqu’à un point maximum, puis diminue.

Q24. Quelle métrique privilégier pour comparer un modèle linéaire et une forêt aléatoire sur les mêmes données de test?

A) R^2 d’entraînement uniquement

B) RMSE ou MAE sur les données de test

C) Nombre de paramètres

D) Temps d’entraînement

Réponse: B) RMSE ou MAE sur les données de test

Pour comparer des modèles de nature différente, on utilise des métriques de performance sur données de test (out-of-sample). RMSE et MAE sont comparables entre modèles, contrairement au R^2 qui peut être biaisé.

Q25. Dans le contexte réglementaire bancaire (Bâle), pourquoi la régression logistique est-elle souvent préférée aux modèles “boîte noire” pour le scoring?

A) Elle est toujours plus performante

B) Elle offre une meilleure explicabilité et interprétabilité des coefficients

C) Elle n’a pas besoin de données

D) Elle est plus rapide à entraîner

Réponse: B) Elle offre une meilleure explicabilité et interprétabilité des coefficients

Les régulateurs (Bâle, BRH) exigent que les modèles de scoring soient explicables. Les coefficients de la régression logistique (via odds ratios) permettent d’expliquer pourquoi un crédit est refusé - exigence légale du “droit à l’explication”.

Résumé des Concepts Clés

Diagnostics Avancés

- **VIF > 10:** Multicolinéarité sévère
- **DW ≠ 2:** Autocorrélation (< 2 positive, > 2 négative)
- **Breusch-Pagan p < 0.05:** Hétéroscédasticité
- **Shapiro-Wilk p > 0.05:** Normalité OK

Régression Logistique

- **OR = e^β:** Interprétation multiplicative
- **Pseudo-R^2 0.2-0.4:** Bon ajustement (McFadden)
- **Logit(p) = log(p/(1-p)):** Fonction de lien

Régularisation

- **Ridge (L2):** Shrink vers 0, jamais exactement 0
- **Lasso (L1):** Peut mettre coefficients = 0 (sélection)
- **λ plus grand:** Plus de pénalisation

Score: ___/25