

Test de Préparation: Régression Linéaire et Diagnostics

Informations

- **Durée estimée:** 45 minutes
 - **Nombre de questions:** 25
 - **Niveau:** Intermédiaire-Avancé
 - **Thèmes:** Régression simple et multiple, hypothèses LINE, diagnostics, interprétation
-

Section 1: Concepts Fondamentaux (5 questions)

Question 1

Dans le modèle de régression $Y = \beta_0 + \beta_1X + \epsilon$, que représente β_1 ?

- A) L'ordonnée à l'origine
- B) Le changement moyen de Y pour une unité de changement de X
- C) L'erreur de prédiction
- D) La variance de Y

Réponse

B) Le changement moyen de Y pour une unité de changement de X

β_1 est la pente de la droite de régression. Elle indique de combien Y change en moyenne lorsque X augmente d'une unité, toutes autres choses égales par ailleurs.

Mnémotechnique: “ β_1 = Bouvement de Y pour 1 unité de X”

Question 2

Une banque souhaite prédire le montant de prêt accordé (Y) en fonction du revenu (X). Le coefficient $\beta_1 = 0.25$. Comment interpréter ce résultat?

- A) 25% des clients obtiennent un prêt
- B) Pour chaque gourde de revenu supplémentaire, le prêt augmente de 0.25 HTG
- C) Le revenu explique 25% de la variance du prêt
- D) La corrélation est de 0.25

Réponse

B) Pour chaque gourde de revenu supplémentaire, le prêt augmente de 0.25 HTG

Le coefficient de régression s'interprète comme l'effet marginal de X sur Y. Ici, chaque augmentation d'1 HTG de revenu est associée à une augmentation de 0.25 HTG du montant de prêt accordé.

Note bancaire: On pourrait aussi dire que pour chaque 10,000 HTG de revenu supplémentaire, le prêt augmente de 2,500 HTG.

Question 3

Quelle est la méthode utilisée pour estimer les coefficients de la régression linéaire?

- A) Maximum de vraisemblance

- B) Gradient descent
- C) Moindres Carrés Ordinaires (MCO/OLS)
- D) Méthode de Newton-Raphson

Réponse

C) Moindres Carrés Ordinaires (MCO/OLS)

La méthode OLS (Ordinary Least Squares) minimise la somme des carrés des résidus: $\sum(Y_i - \hat{Y}_i)^2$

C'est la méthode standard pour la régression linéaire car elle fournit les estimateurs BLUE (Best Linear Unbiased Estimators) sous les hypothèses classiques.

Question 4

Quelle est la différence entre R^2 et R^2 ajusté?

- A) R^2 ajusté est toujours plus grand que R^2
- B) R^2 ajusté pénalise l'ajout de variables non significatives
- C) R^2 ajusté ne s'applique qu'à la régression simple
- D) Il n'y a pas de différence significative

Réponse

B) R^2 ajusté pénalise l'ajout de variables non significatives

Formule: R^2 ajusté = $1 - [(1 - R^2)(n - 1) / (n - p - 1)]$

- R^2 augmente TOUJOURS quand on ajoute une variable
- R^2 ajusté peut DIMINUER si la variable ajoutée n'améliore pas suffisamment le modèle
- Utiliser R^2 ajusté pour comparer des modèles avec différents nombres de variables

Mnémotechnique: "R² ajusté = R² avec Ajustement pour le nombre de prédicteurs"

Question 5

Dans une régression avec $R^2 = 0.75$, que signifie cette valeur?

- A) 75% des observations sont bien classées
- B) 75% de la variance de Y est expliquée par le modèle
- C) Le coefficient de corrélation est 0.75
- D) L'erreur de prédiction est de 75%

Réponse

B) 75% de la variance de Y est expliquée par le modèle

R^2 (coefficient de détermination) mesure la proportion de la variance totale de Y qui est expliquée par les variables X du modèle.

- $R^2 = 0$: Le modèle n'explique rien
 - $R^2 = 1$: Le modèle explique parfaitement Y
 - $R^2 = 0.75$: 75% expliqué, 25% non expliqué (résiduel)
-

Section 2: Hypothèses LINE (5 questions)

Question 6

Que signifie l'acronyme LINE dans le contexte des hypothèses de la régression?

- A) Linéarité, Indépendance, Normalité, Égalité des variances
- B) Logarithme, Itération, Numérique, Erreur
- C) Limite, Intervalle, Nombre, Estimation
- D) Lien, Impact, Niveau, Effet

Réponse

A) Linéarité, Indépendance, Normalité, Égalité des variances

Les quatre hypothèses fondamentales de la régression linéaire: - **Linéarité**: Relation linéaire entre X et Y - **Indépendance**: Les observations sont indépendantes - **Normalité**: Les résidus suivent une loi normale - **Égalité des variances (Homoscédasticité)**: Variance constante des résidus

Question 7

Comment s'appelle le problème lorsque la variance des résidus n'est pas constante?

- A) Multicolinéarité
- B) Autocorrélation
- C) Hétéroscédasticité
- D) Endogénéité

Réponse

C) Hétéroscédasticité

- Homo-scédasticité: Variance constante (ce qu'on veut)
- Hétéro-scédasticité: Variance non constante (problème)

Détection: - Graphique résidus vs valeurs ajustées (forme d'entonnoir = problème) - Test de Breusch-Pagan - Test de White

Solutions: - Transformation log de Y - Régression avec erreurs robustes (HC3) - Régression pondérée (WLS)

Question 8

Le test de Durbin-Watson donne une valeur de 0.8. Que conclure?

- A) Pas de problème, le modèle est bon
- B) Il y a une autocorrélation positive des résidus
- C) Il y a une autocorrélation négative des résidus
- D) Les résidus ne sont pas normaux

Réponse

B) Il y a une autocorrélation positive des résidus

Interprétation du test de Durbin-Watson: - $DW \approx 2$: Pas d'autocorrélation (idéal) - $DW < 1.5$: Autocorrélation positive (problème) - $DW > 2.5$: Autocorrélation négative (problème)

Une valeur de 0.8 est bien en dessous de 1.5, indiquant une autocorrélation positive forte.

Action: Inclure des variables lag, ou utiliser des modèles pour données autocorrélées (GLS, ARIMA).

Question 9

Quel test utilisez-vous pour vérifier la normalité des résidus?

- A) Test de Durbin-Watson
- B) Test de Breusch-Pagan
- C) Test de Shapiro-Wilk
- D) Test de VIF

Réponse

C) Test de Shapiro-Wilk

Tests de normalité: - **Shapiro-Wilk:** Pour $n < 5000$ (le plus puissant) - **Jarque-Bera:** Basé sur asymétrie et kurtosis - **Kolmogorov-Smirnov:** Alternative

Interprétation: - $p > 0.05 \rightarrow$ Résidus normaux (OK) - $p < 0.05 \rightarrow$ Résidus non normaux (problème)

Visualisation: QQ-plot (les points doivent suivre la diagonale)

Question 10

Vous observez un pattern en forme de U dans le graphique des résidus vs valeurs ajustées. Que cela suggère-t-il?

- A) Hétéroscédasticité
- B) Violation de la linéarité
- C) Autocorrélation
- D) Multicolinéarité

Réponse

B) Violation de la linéarité

Un pattern systématique (U, courbe) dans le graphique résidus vs fitted indique que la relation n'est pas linéaire.

Solutions: - Ajouter des termes quadratiques (X^2) - Transformer X (log, racine carrée) - Utiliser un modèle non-linéaire - Ajouter des variables manquantes

À retenir: Les résidus doivent être aléatoirement dispersés autour de 0.

Section 3: Multicolinéarité (5 questions)

Question 11

Qu'est-ce que le VIF (Variance Inflation Factor)?

- A) Un test de normalité
- B) Une mesure de la multicolinéarité entre variables explicatives
- C) Un indicateur de la variance de Y
- D) Le rapport entre variance expliquée et non expliquée

Réponse

B) Une mesure de la multicolinéarité entre variables explicatives

$$VIF = 1 / (1 - R^2_j)$$

Où R^2_j est le R^2 de la régression de X_j sur toutes les autres variables X.

Interprétation: - VIF = 1: Pas de corrélation avec les autres X - VIF = 1-5: Corrélation modérée (acceptable) - VIF = 5-10: Corrélation élevée (attention) - VIF > 10: Multicolinéarité sévère (action requise)

Question 12

Une variable a un VIF de 15. Quelle action recommandez-vous?

- A) Aucune action, c'est acceptable
- B) Supprimer la variable ou combiner avec une autre variable corrélée
- C) Ajouter plus de variables au modèle
- D) Augmenter la taille de l'échantillon

Réponse

B) Supprimer la variable ou combiner avec une autre variable corrélée

Un VIF de 15 indique une multicolinéarité sévère. Solutions: 1. Supprimer une des variables corrélées 2. Combiner les variables en un indice 3. Utiliser l'ACP pour réduire les dimensions 4. Régression Ridge (pénalisation L2)

Conséquences si non traité: - Coefficients instables - Intervalles de confiance très larges - Difficultés d'interprétation

Question 13

Deux variables (revenu et patrimoine) ont une corrélation de 0.85. Comment gérer cette situation dans une régression?

- A) Ignorer, ce n'est pas un problème
- B) Les inclure toutes les deux sans modification
- C) Créer un indice composite ou n'en garder qu'une
- D) Ajouter une troisième variable

Réponse

C) Créer un indice composite ou n'en garder qu'une

Une corrélation de 0.85 est très élevée et causera de la multicolinéarité.

Options: 1. Ne garder que la variable la plus pertinente théoriquement 2. Créer un indice: richesse = (revenu + patrimoine) / 2 3. Utiliser l'ACP pour extraire un facteur commun 4. Appliquer une régression Ridge

Règle pratique: Éviter les paires de variables avec $r > 0.7$

Question 14

Comment la multicolinéarité affecte-t-elle le pouvoir prédictif du modèle?

- A) Elle le détruit complètement
- B) Elle n'affecte pas les prédictions, seulement l'interprétation des coefficients
- C) Elle améliore les prédictions
- D) Elle augmente le R²

Réponse

B) Elle n'affecte pas les prédictions, seulement l'interprétation des coefficients

Paradoxe de la multicolinéarité: - Le modèle peut toujours bien PRÉDIRE - Mais on ne peut pas interpréter les coefficients individuellement - Les coefficients sont instables (haute variance) - Les tests t individuels ne sont pas fiables

Quand c'est un problème: Si l'objectif est d'interpréter l'effet de chaque variable. **Quand c'est moins grave:** Si l'objectif est uniquement la prédiction.

Question 15

Quel code Python permet de calculer le VIF pour toutes les variables?

- A) model.vif()
- B) variance_inflation_factor() de statsmodels
- C) df.corr()
- D) model.rsquared

Réponse

B) variance_inflation_factor() de statsmodels

```
from statsmodels.stats.outliers_influence import variance_inflation_factor
import pandas as pd

def calculate_vif(X):
    vif_data = pd.DataFrame()
    vif_data["Variable"] = X.columns
    vif_data["VIF"] = [variance_inflation_factor(X.values, i)
                      for i in range(X.shape[1])]
    return vif_data.sort_values('VIF', ascending=False)

vif_results = calculate_vif(X)
print(vif_results)
```

Section 4: Tests d'Hypothèses (5 questions)

Question 16

Le F-test global dans une régression teste:

- A) Si chaque coefficient est significatif
- B) Si au moins un coefficient est différent de zéro
- C) Si les résidus sont normaux
- D) Si la variance est constante

Réponse

B) Si au moins un coefficient est différent de zéro

Le F-test global: - $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ (aucune variable n'est utile) - $H_1: \text{Au moins un } \beta_j \neq 0$ (au moins une variable est utile)

$$F = (\text{SSR}/p) / (\text{SSE}/(n-p-1)) = \text{MSR}/\text{MSE}$$

Si $\text{Prob}(F) < 0.05 \rightarrow$ Le modèle dans son ensemble est significatif.

Attention: Un F-test significatif ne dit pas QUELLES variables sont significatives.

Question 17

Dans le summary d'un modèle, la variable "age" a un coefficient de 500 avec une p-value de 0.23. Que conclure?

- A) L'âge a un effet significatif sur Y
- B) L'âge n'a pas d'effet statistiquement significatif sur Y
- C) L'âge devrait être transformé
- D) Le modèle est mal spécifié

Réponse

B) L'âge n'a pas d'effet statistiquement significatif sur Y

Interprétation de la p-value pour un coefficient: - $p < 0.05 \rightarrow$ Coefficient significatif ($H_0: \beta = 0$ rejetée) - $p \geq 0.05 \rightarrow$ Coefficient non significatif (H_0 non rejetée)

Avec $p = 0.23 > 0.05$, on ne peut pas conclure que l'âge a un effet sur Y.

Action possible: Considérer retirer cette variable du modèle (surtout si le R^2 ajusté augmente).

Question 18

L'intervalle de confiance à 95% pour β_1 est [0.15, 0.45]. Comment interpréter?

- A) β_1 n'est pas significatif
- B) β_1 est significatif et positif
- C) Il y a 95% de chances que $\beta_1 = 0.30$
- D) L'erreur standard est de 0.15

Réponse

B) β_1 est significatif et positif

Si l'IC à 95% ne contient pas 0, alors β_1 est statistiquement significatif à $\alpha = 0.05$.

Ici, [0.15, 0.45] ne contient pas 0, donc: - β_1 est significativement différent de 0 - β_1 est positif (effet positif de X sur Y) - On est 95% confiant que la vraie valeur de β_1 est entre 0.15 et 0.45

Mnémotechnique: "Si 0 n'est pas dans l'IC, le coefficient est OK (significatif)"

Question 19

Comment interpréter un t-statistic de -4.5 pour un coefficient?

- A) Le coefficient est négatif et non significatif
- B) Le coefficient est négatif et hautement significatif
- C) Il y a une erreur dans le modèle
- D) Le coefficient doit être transformé

Réponse

B) Le coefficient est négatif et hautement significatif

Le t-statistic = $\hat{\beta} / \text{SE}(\hat{\beta})$

- Le signe indique la direction de l'effet (négatif ici)
- La magnitude indique la force de l'évidence
- $|t| > 2 \rightarrow$ Généralement significatif
- $|t| = 4.5 \rightarrow$ Très significatif (p-value très faible)

Règle rapide: $|t| > 1.96 \rightarrow$ significatif à 5%

Question 20

Quelle métrique utiliser pour comparer deux modèles avec des nombres différents de variables?

- A) R^2
- B) R^2 ajusté ou AIC/BIC
- C) F-statistic
- D) t-statistic

Réponse

B) R^2 ajusté ou AIC/BIC

Pour comparer des modèles: - **R^2 ajusté:** Pénalise l'ajout de variables inutiles - **AIC (Akaike):** $AIC = 2k - 2\ln(L)$, plus bas = meilleur - **BIC (Bayesian):** $BIC = k \times \ln(n) - 2\ln(L)$, plus bas = meilleur

Ne PAS utiliser R^2 car il augmente toujours avec plus de variables.

Règle: Choisir le modèle avec le plus bas AIC/BIC ou le plus haut R^2 ajusté.

Section 5: Applications Bancaires (5 questions)

Question 21

Une banque modélise le montant de défaut comme: $\text{Défaut} = 50000 - 0.02 \times \text{Revenu} + 15000 \times \text{Ratio_Endettement} + \varepsilon$

Si un client a un revenu de 1,000,000 HTG et un ratio d'endettement de 0.4, quel est le défaut prédict?

- A) 30,000 HTG
- B) 36,000 HTG
- C) 56,000 HTG
- D) 76,000 HTG

Réponse

B) 36,000 HTG

Calcul: Défaut = $50000 - 0.02 \times (1,000,000) + 15000 \times (0.4)$ Défaut = $50000 - 20000 + 6000$
Défaut = 36,000 HTG

Interprétation des coefficients: - Revenu: Effet négatif (-0.02), un revenu plus élevé réduit le défaut - Ratio endettement: Effet positif (+15000), plus d'endettement augmente le défaut

Question 22

Dans un modèle de scoring crédit, pourquoi préférer la régression logistique à la régression linéaire?

- A) La régression logistique est plus rapide
- B) La variable cible (défaut oui/non) est binaire, pas continue
- C) La régression logistique a un meilleur R²
- D) La régression linéaire ne fonctionne pas avec des données bancaires

Réponse

B) La variable cible (défaut oui/non) est binaire, pas continue

Différences clés: | Aspect | Régression Linéaire | Régression Logistique | |-----|-----|-----|-----| | Y | Continue | Binaire (0/1) | | Prédiction | Valeur numérique | Probabilité (0-1) | | Fonction | $Y = \beta X$ | $P(Y=1) = 1/(1+e^{(-\beta X)})$ | | Métrique | R², RMSE | AUC, Gini |

Pour le scoring crédit (défaut = 0 ou 1), la régression logistique est appropriée.

Question 23

Un analyste obtient un R² de 0.98 sur ses données d'entraînement mais seulement 0.45 sur les données de test. Quel est le problème probable?

- A) Les données de test sont incorrectes
- B) Le modèle est en underfitting
- C) Le modèle est en overfitting
- D) Le R² est mal calculé

Réponse

C) Le modèle est en overfitting

L'overfitting (surapprentissage) se caractérise par: - Excellente performance sur les données d'entraînement - Mauvaise performance sur les nouvelles données

Causes possibles: - Trop de variables par rapport aux observations - Modèle trop complexe - Pas de régularisation

Solutions: - Réduire le nombre de variables - Utiliser la validation croisée - Appliquer une régularisation (Ridge, Lasso) - Augmenter la taille de l'échantillon

Question 24

Comment interpréter un intervalle de prédiction plus large qu'un intervalle de confiance?

- A) C'est une erreur de calcul
- B) L'intervalle de prédiction inclut l'incertitude sur Y individuel (inclut ϵ)
- C) Le modèle est mal spécifié
- D) Il faut plus de données

Réponse

B) L'intervalle de prédiction inclut l'incertitude sur Y individuel (inclut ϵ)

Deux types d'intervalles: - **Intervalle de Confiance (IC):** Incertitude sur $E[Y|X]$ (moyenne prédictive) - **Intervalle de Prédiction (IP):** Incertitude sur Y individuel

$IP = IC + \text{variance du terme d'erreur } \epsilon$

L'IP est TOUJOURS plus large que l'IC car il inclut: 1. L'incertitude sur les paramètres estimés (comme l'IC) 2. L'incertitude sur la valeur individuelle (variance de ϵ)

Question 25

Une banque utilise un modèle de régression pour estimer le solde client. Quelles variables seraient logiquement des prédicteurs pertinents?

- A) Couleur des yeux, taille, signe astrologique
- B) Revenu, ancienneté, nombre de produits détenus, segment
- C) Date de naissance uniquement
- D) Numéro de compte uniquement

Réponse

B) Revenu, ancienneté, nombre de produits détenus, segment

Variables pertinentes pour prédire le solde client: - **Revenu:** Plus de revenu = capacité d'épargne plus élevée - **Ancienneté:** Clients fidèles accumulent plus - **Nb produits:** Engagement plus fort avec la banque - **Segment:** Premium/VIP ont généralement des soldes plus élevés - **Âge:** Cycle de vie (accumulation vs décumulation)

À éviter: Variables sans lien logique avec Y (éviter le data dredging)

Résumé et Mnémotechniques

“RHINO” - Étapes de Diagnostic

- R - R^2 et R^2 ajusté: Pouvoir explicatif
- H - Hypothèses LINE: Vérifier toutes les 4
- I - Intervalles: IC et IP pour prédictions
- N - Normalité: Résidus normaux (Shapiro-Wilk)
- O - Outliers: Points influents (Cook's distance)

“VISC” - Traiter la Multicolinéarité

- V - VIF: Calculer pour chaque variable
- I - Identifier: Corrélations > 0.7

S - Supprimer: Variable redondante
C - Combiner: Créer un indice

“CIPT” - Interpréter les Coefficients

C - Coefficient: Magnitude et signe
I - Intervalle: IC à 95%
P - P-value: < 0.05 = significatif
T - T-statistic: Force de l'évidence

Seuils à Retenir

Métrique	Seuil	Interprétation
P-value	< 0.05	Significatif
VIF	$> 5-10$	Problème de multicolinéarité
Durbin-Watson	1.5-2.5	Acceptable
R ²	> 0.3	Acceptable pour données réelles

Score et Auto-évaluation

Score	Niveau	Recommandation
23-25	Expert	Prêt pour l'examen
18-22	Avancé	Réviser les points faibles
13-17	Intermédiaire	Revoir le document complet
< 13	Débutant	Étude approfondie nécessaire

Test préparé pour l'examen Data Analyst - UniBank Haiti Thème: Régression Linéaire - L'outil fondamental