MATH60604
Modélisation statistique
Tests d'hypothèses pour modèles linéaires

Léo Belzile

HEC Montréal
Département de sciences de la décision

Résumé des postulats du modèle linéaire

1. **Indépendance**: les erreurs $\varepsilon_1, \ldots, \varepsilon_n$ sont des variables aléatoires indépendantes sachant $\mathbf{X}$.
   - Cela implique que Y_i est conditionellement indépendant des autres réponses sachant $\mathbf{X}_i$.
2. **Homoscédasticité**: la variance des erreurs est constante, soit $\text{Var}(\varepsilon_i) = \sigma^2$ pour $i = 1, \ldots, n$. Si la variance n'est pas constante, les erreurs sont dites par opposition hétéroscédastiques.
3. **Normalité**: les erreurs, ε, suivent une loi normale.
   - cette supposition est uniquement requise pour la validité des tests d'hypothèse et des intervalles de confiance.

Résumé des postulats du modèle linéaire

4. **Linéarité**: la surface de réponse du modèle linéaire,

$$\mathsf{E}\left(Y \mid \mathsf{X}_1, \ldots, \mathsf{X}_p\right) = \beta_0 + \beta_1 \mathsf{X}_1 + \ldots + \beta_p \mathsf{X}_p.$$

est correctement spécifiée.

- L'espérance des erreurs est zéro, $\mathsf{E}\left(\varepsilon_i \mid \mathbf{X}_i\right) = 0$ pour $i = 1, \ldots, n$.
- Toutes les variables explicatives importantes sont incluses dans le modèle.
- Leurs effets (supposés linéaires en β) sont correctement modélisés.

Tests d'hypothèses pour paramètres individuels

- En inférence, il est souvent important de tester si l'effet d'une variable explicative est statistiquement significative.
- Cela revient à tester si le coefficient associé à la variable est différent de zéro,

$$\mathscr{H}_0 : \beta_j = 0, \qquad \mathscr{H}_1 : \beta_j \neq 0$$

- Pour tester cette hypothèse bilatérale, on utilise une statistique de Wald,

$$t = \frac{\widehat{\beta}_j - 0}{\mathsf{se}(\widehat{\beta}_j)},$$

où $\mathsf{se}(\widehat{\beta}_j)$ est l'estimé de l'**erreur-type** de $\widehat{\beta}_j$. Sa distribution nulle est Student-t et est rapportée par SAS sous le vocable **Valeur du test t**.

Est-ce que les gens sont prêts à payer plus par carte de crédit?

On va travailler avec l'exemple suivant inspiré de l'article

Référence

Prelec, D. et Simester, D. (2001). Always Leave Home Without It: A Further Investigation of the Credit-Card Effect on Willingness to Pay. *Marketing Letters*, **12**, 5-12.

- **Question de l'étude**: Est-ce que le fait de payer par carte de crédit encourage les gens à payer plus?
- **Contexte**: Lors de transactions de valeurs potentiellement élevées, le montant offert par les gens devant payer par carte de crédit peut être plus élevé que celui des gens devant payer en billets.
- **Objectifs**: Présenter de nouveaux éléments en faveur de la proposition à l'effet que les individus sont enclins à payer plus pour un produit lorsqu'ils utilisent une carte de crédit.

Exemple: carte de crédit vs comptant

- Le produit en vente: paire de billets pour le dernier match de la saison des Celtics de Boston (équipe de basketball de la NBA). Cette partie était importante, car elle allait décider qui allait terminer premier dans cette division.
- 64 sujets répartis **au hasard** dans deux groupes:
  - Groupe 1 (33 sujets) : individus doivent payer comptant
  - Groupe 2 (31 sujets): individus doivent payer par carte de crédit
- Tous les sujets ont dû remplir un questionnaire où on leur demandait combien ils étaient prêts à payer pour la paire de billets.

Allocation aléatoire

- Assigner les sujets au hasard aux différentes conditions expérimentales (groupes) s'appelle la randomisation.
- **Idée**: tenter d'avoir des groupes qui sont le plus semblables possible quant aux caractéristiques qui pourraient influencer la variable réponse et, ainsi, diminuer leur impact.
  - il se pourrait que l'âge, le sexe, ainsi que d'autres variables, aient une influence sur le montant d'argent que la personne est prête à offrir.
- Autre manière de procéder: contrôler directement pour les effets de ces variables dans un modèle de régression linéaire.

Exemple: moyen de paiement

La base de données `billets` contient des données simulées correspondant à cette étude, incluant les variables

- `offre`: montant offert (en $) par l'individu pour se procurer les billets
- `groupe`: identifie le groupe d'appartenance
  - `groupe=0` pour les individus devant payer comptant
  - `groupe=1` pour les individus devant payer avec une carte de crédit

Objectif général

Comparer une **moyenne** (montant moyen d'argent que les gens sont prêts à payer) entre deux groupes (carte de crédit versus comptant)

Test-t pour deux échantillons avec un modèle de régression

- La variable `offre` représente le montant en dollars et la variable `groupe` l'indicateur binaire (0 pour comptant, 1 pour crédit).
- Le test d'égalité des moyennes est équivalent à tester si l'effet de `groupe` sur l'`offre` est nul.
- On peut formuler ce problème à l'aide du modèle de régression

$$\texttt{offre} = \beta_0 + \beta_1 \texttt{groupe} + \varepsilon$$

- Tester $\mathscr{H}_0 : \mu_{\texttt{comptant}} = \mu_{\texttt{crédit}}$ revient à tester $\mathscr{H}_0 : \beta_1 = 0$.

Régression linéaire avec données `billets`

Paramètre	Estimation	Erreur type	Valeur du test t	Pr > |t|
Constante	56.60606061	3.00335700	18.85	<.0001
groupe	15.00684262	4.31535073	3.48	0.0009

- Nous estimons que la différence **moyenne** entre les deux groupes est $\widehat{\beta}_1 = \$15$, soit des offres par crédits plus élevées.
- Cette différence est statistiquement significatif à niveau 5% (valeur-p de 0.0009).
- La sortie du modèle de régression linéaire ne contient que la valeur-p du test bilatéral, mais celle du test unilatéral est moitié moindre (à cause de la symmétrie de la loi Student-t).

Intervalles de confiance

Dans SAS, la commande `clparm` ajoute les intervalles de confiance de Wald de niveau $(1-\alpha)$ (par défaut 95%) au tableau des estimés.

- Le code qui suit cette procédure pour le modèle régression simple

$$\texttt{intention}_i = \beta_0 + \beta_1 \texttt{fixation}_i + \varepsilon_i.$$

Code SAS pour ajuster le modèle linéaire

```
proc glm data=infe.intention;
model intention=fixation / ss3 solution clparm;
run;
```

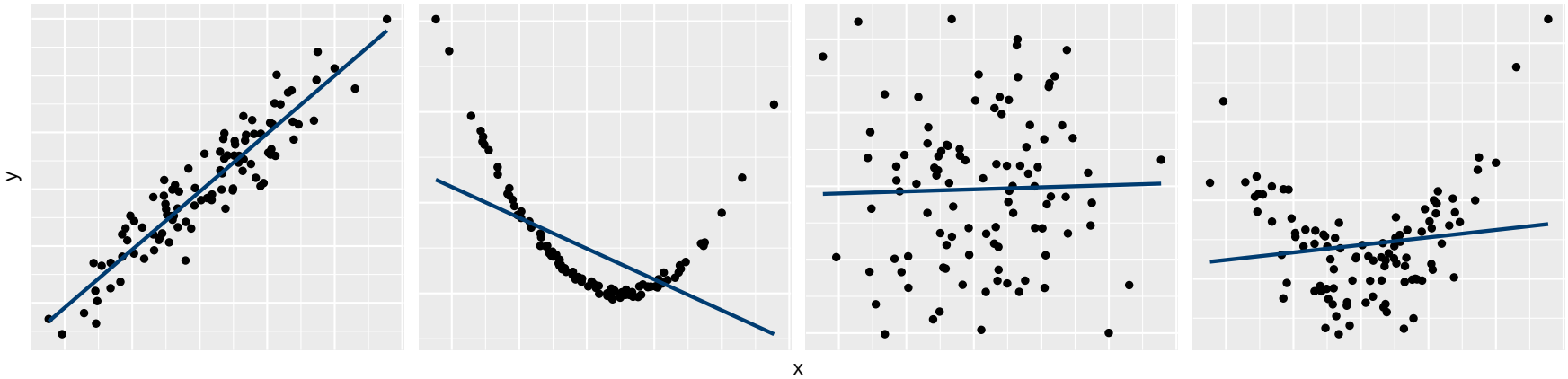
Tests pour paramètres individuels et intervalles de confiance

Paramètre	Estimation	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à95%	
Constante	6.453188456	0.42849218	15.06	<.0001	5.604657280	7.301719632
fixation	1.144083751	0.22351452	5.12	<.0001	0.701464150	1.586703353

- La valeur de la statistique-t pour $\mathscr{H}_0 : \beta_1 = 0$ versus $\mathscr{H}_1 : \beta_1 \neq 0$ est $t = 1.144/0.224 = 5.12$.
- La valeur-p du test bilatéral est inférieure à 0.0001.
- On rejette $\mathscr{H}_0$, car il y a un effet significatif du temps de `fixation` sur l'`intention` d'achat.
- L'intervalle de confiance à 95% pour β_1, correspondant à l'effet linéaire de `fixation`, est $[0.70; 1.59]$.
- Puisque l'intervalle ne contient pas 0, on déduit que le paramètre est significativement différent de zéro à niveau $\alpha = 5\%$.

Interprétation du test d'hypothèse

Si on rejette $\mathscr{H}_0 : \beta_j = 0$ pour l'alternative, on exclut la possibilité qu'il n'y ait aucune **relation linéaire** significative entre X_j et Y **une fois l'effet des autres variables pris en compte**.



On rejette $\mathscr{H}_0 : \beta_1 = 0$ dans les deux graphiques de gauche (valeurs-p moins de 10^{-6}, mais pas dans ceux de droite (valeurs-p de 0.787 et 0.156). Le coefficient de détermination est, de gauche à droite, 0.87, 0.3, 10^{-3} et 10^{-3}.

Tests d'hypothèses pour plusieurs paramètres

- Il est également possible de tester si plusieurs β sont simultanément zéros, par exemple

$$\mathscr{H}_0 : \beta_{\texttt{age}} = \beta_{\texttt{emotion}} = 0.$$

- Pour valider l'utilité d'une variable catégorielle, il convient également de tester pour son effet global.
  - La variable `educ` a trois niveaux et est modélisée à l'aide de deux variables indicatrices `educ1` et `educ2`; tester l'effet global d'`educ` revient à tester

$$\mathscr{H}_0 : \beta_{\texttt{educ}_1} = \beta_{\texttt{educ}_2} = 0.$$

Comparaison de modèles emboîtés

- Soit le modèle complet contenant p régresseurs,

$$\mathbb{M}_1 : Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_g X_g + \beta_{k+1} X_{k+1} + \ldots + \beta_p X_p + \varepsilon.$$

- On considère le test

$$\mathscr{H}_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p = 0,$$

- soit l'hypothèse que $(p - k)$ des β (sans perte de généralité les derniers) sont simultanément zéros.
- Le modèle contraint contient seulement les covariables pour lesquelles $\beta_j \neq 0$,

$$\mathbb{M}_0 : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k + \varepsilon.$$

Tests-F pour comparaison de modèles imbriqués (optionnel)

- Soit $\mathrm{SS}_e(\mathbb{M}_1)$ la somme du carré des résidus du modèle complet $\mathbb{M}_1$,

$$\mathrm{SS}_e(\mathbb{M}_1) = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i^{\mathbb{M}_1})^2,$$

  où $\widehat{Y}_i^{\mathbb{M}_1}$ est la ie valeur ajustée du modèle $\mathbb{M}_1$.
- On définie de la même façon la somme du carré des résidus, $\mathrm{SS}_e(\mathbb{M}_0)$, pour le modèle $\mathbb{M}_0$.
- Logiquement, $\mathrm{SS}_e(\mathbb{M}_0) \geq \mathrm{SS}_e(\mathbb{M}_1)$ (pourquoi?)

La statistique du test-F est

$$F = \frac{\{\mathrm{SS}_e(\mathbb{M}_0) - \mathrm{SS}_e(\mathbb{M}_1)\}/(p-k)}{\mathrm{SS}_e(\mathbb{M}_1)/(n-p-1)}$$

- Sous $\mathscr{H}_0$, la statistique-F suit une loi de Fisher avec $(p-k)$ et $(n-p-1)$ degrés de liberté, $\mathrm{F}(p-k, n-p-1)$.

$p-k$ est le nombre de restrictions, $n-p-1$ est la taille de l'échantillons moins le nombre de β de $\mathbb{M}_1$.

Test-F pour l'effet de toutes les covariables

- Le premier tableau de la sortie SAS avec la procédure `glm` donne le résultat du test global pour tous les coefficients $\mathscr{H}_0 : \beta_1 = \ldots = \beta_p = 0$ contre l'hypothèse alternative qu'au moins un des paramètres est utile pour prédire Y et que son effet est non-nul.
- Dans le modèle qui inclut toutes les variables, cela revient à tester $\mathscr{H}_0 : \beta_{\texttt{sexe}} = \beta_{\texttt{age}} = \cdots = \beta_{\texttt{emotion}} = 0$.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	9	460.965432	51.218381	9.99	<.0001
Erreur	110	564.026235	5.127511		
Total sommes corrigées	119	1024.991667			

- On rejette $\mathscr{H}_0$, ce qui veut dire que l'effet linéaire d'au moins un régresseur est non-nul pour décrire l'intention d'achat.

Décomposition de la somme des carrés

- Les comparaisons de modèles emboîtés sont effectuées avec la somme des carrés de **type III**.
  - le modèle contraint inclut toutes les variables explicatives, hormis celle testée.
- La somme des carrés de type I fait des **comparaisons séquentielles**. Pour la *j*e covariable, on compare le modèle avec les variables préalablements ajoutées — l'ordre d'entrée des variables explicatives devient important.
  - par exemple, si la première variable ajoutée est celle dont vous voulez tester l'effet, la comparaison est entre un modèle avec seulement cette variable et le modèle avec uniquement l'ordonnée à l'origine.
- Pour éviter de vous tromper dans l'interprétation, ajoutez `ss3` dans votre appel SAS.

Test-F global pour variables catégorielles

- Quand la je variable explicative est continue ou binaire, le test-F est équivalent au test-t pour $\beta_j = 0$. Ces variables ont `DDL=1`.
- Quand les variables sont catégorielles (telles que définies par `class` en SAS), la sortie est différente. Par exemple, le test global pour une variable catégorielle `cat` à k niveaux correspond à l'hypothèse $\mathscr{H}_0 : \beta_{\texttt{cat}_1} = \cdots = \beta_{\texttt{cat}_{k-1}} = 0$
- comparer le modèle avec ou sans cette variable catégorielle, en prenant en compte l'effet de toutes les autres variables.

Supposez qu'on ajuste un modèle linéaire avec toutes les variables explicatives des données `intention`.

Code SAS pour ajuste le modèle linéaire complet

```
proc glm data=modstat.intention noprint;
class sexe educ revenu;
model intention= fixation emotion
    sexe age revenu educ statut / ss3 solution;
run;
```

- Les degrés de libertés pour `revenu` et `educ` sont deux, parce que chaque variable a trois catégories.
- par exemple, le test F compare le modèle avec toutes les variables explicatives à celui sans éducation, $\beta_{\texttt{educ}_1} = \beta_{\texttt{educ}_2} = 0$.

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
sexe	1	32.9762187	32.9762187	6.43	0.0126
age	1	37.1972654	37.1972654	7.25	0.0082
revenu	2	49.8556095	24.9278047	4.86	0.0095
educ	2	14.8380494	7.4190247	1.45	0.2397
statut	1	2.2349782	2.2349782	0.44	0.5105
fixation	1	184.7635934	184.7635934	36.03	<.0001
emotion	1	36.9122275	36.9122275	7.20	0.0084