MATH60604
Modélisation statistique
§ 2f Prédiction

Léo Belzile

HEC Montréal
Département de sciences de la décision

Valeurs ajustées et prédictions

- En commercialisation par base de données, le but premier est de développer un modèle pour pouvoir obtenir des prévisions de la variable dépendante et ensuite prendre des décisions d'affaires basées sur elles.
- Par exemple, on pourrait vouloir prévoir le montant d'argent dépensé si on envoie une offre à un client.
- La procédure habituelle consiste alors à envoyer l'offre à un échantillon de clients, à développer un modèle avec des données, et ensuite à appliquer ce modèle (obtenir les prédictions) pour les autres clients dans la base de données.

Prédiction

- On veut estimer la moyenne de Y quand $\mathbf{X} = \boldsymbol{x}$,

$$\mathrm{E}\left(Y \mid \mathbf{X} = \boldsymbol{x}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p.$$

- On pourrait également vouloir prédire la valeur d'une nouvelle valeur de la réponse Y_i quand $\mathbf{X}_i = \boldsymbol{x}$; on rappelle que

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon_i.$$

- Que l'on veuille estimer la moyenne ou prédire la valeur de Y quand $\mathbf{X} = \boldsymbol{x}$, les valeurs prédites et ajustées sont égales aux points sur la "droite" correspondant à $\mathbf{X} = \boldsymbol{x}$,

$$\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_p x_p.$$

Davantage d'incertitude pour les prédictions individuelles

- La moyenne estimée et la prédiction pour un individu sont **identiques** pour la régression linéaire.
  - nous verrons que ce n'est pas le cas dans les modèles mixtes, dans lesquels on inclut des effets aléatoires pour des individus ou des groupes.
- Une fois qu'on a obtenu des estimés des paramètres, on peut obtenir les valeurs ajustées et les prédictions pour Y pour un ensemble de variables explicatives $X_1 = x_1, \ldots, X_p = x_p$

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_p x_p.$$

Même prédiction, différentes incertitudes

- Bien que l'estimateur de la moyenne de Y, $\mathrm{E}(Y \mid \mathbf{X})$, soit identique, cet estimateur sera **plus précis** que la valeur prédite d'une nouvelle variable réponse Y_i.
- L'**intervalle de confiance** (pour l'espérance) est plus petit que l'**intervalle de prédiction**.
  - **explication**: une nouvelle observation inclut un terme d'erreur additionnel ε qui augmente l'incertitude.

Valeurs ajustées et intervalles de confiance pour prédictions en SAS

- On crée un nouveau jeu de données `newdata` contenant les valeurs des combinaisons de variables explicatives pour lesquelles on veut obtenir des prédictions.
- Nous utilisons à titre d'exemple la première observation du jeu de données et on copie la combinaison des valeurs explicatives, mais en faisant varier le temps de fixation de 0 à 6 secondes.

Code SAS pour créer un nouveau jeu de données

```
data newdata;
   set infe.intention(obs=1);
   do fixation=0 to 6;
      output;
   end;
run;
```

Sauvegarder la sortie du modèle linéaire

- Par la suite, on ajuste le modèle linéaire et on sauvegarde l'information dans un objet, disons `modelinfo`, afin de faire des prédiction.

Code SAS pour enregistrer la sortie du modèle linéaire

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

Prédiction à partie de la sortie

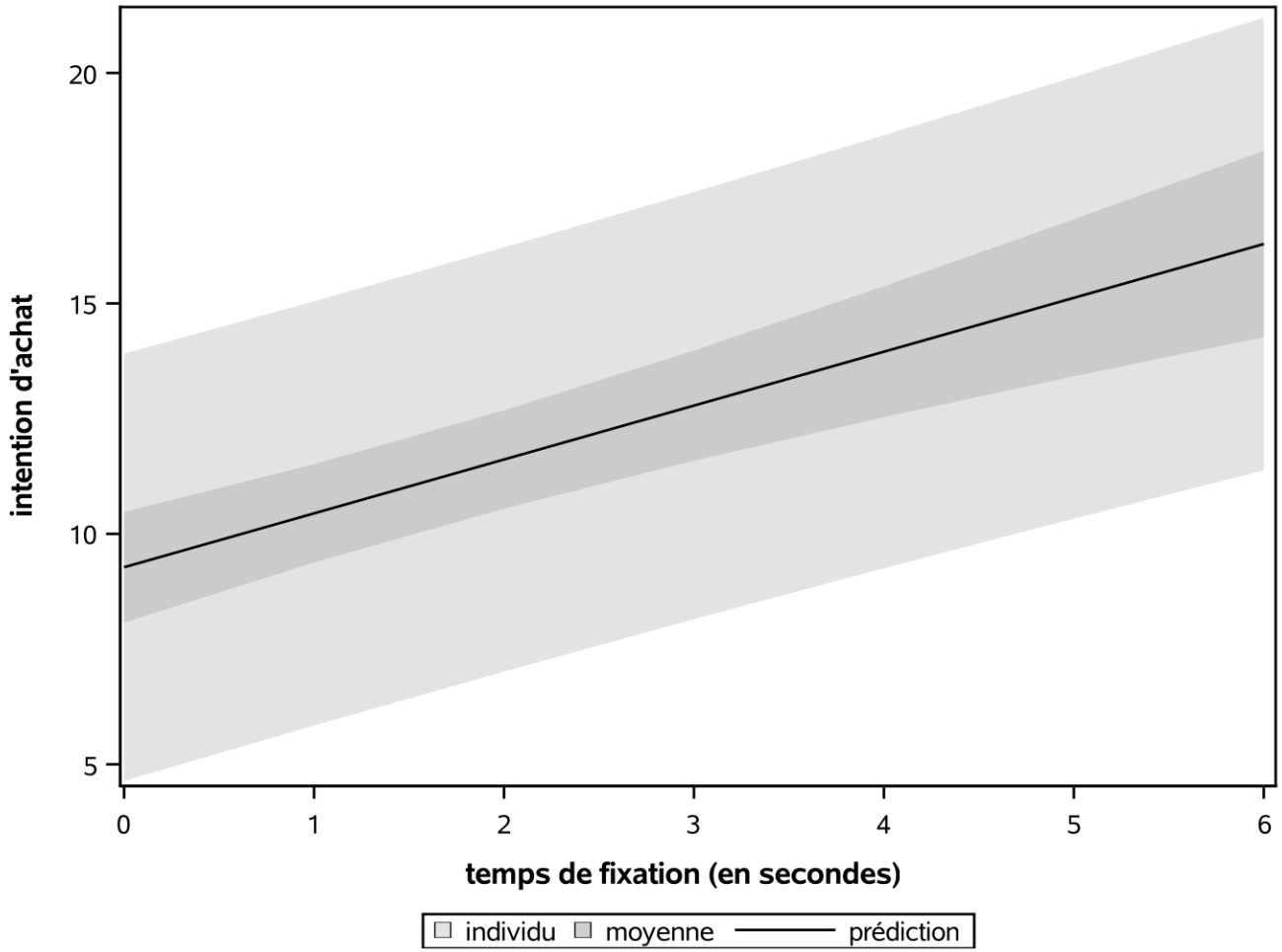
- La procédure `plm` permet d'obtenir les prédictions pour le modèle spécifié par `modelinfo`. Ces prédictions sont enregistrées dans le fichier temporaire `prediction`.
- On inclut l'intervalle de confiance ponctuel pour la moyenne (`lclm` et `uclm`) et l'intervalle de prédiction pour les valeurs individuelles (`lcl` et `ucl`).

Code SAS pour obtenir les prédictions avec `plm`

```
proc plm restore=modelinfo;
score data=newdata out=prediction predicted
    lclm uclm lcl ucl;
run;

proc sgplot data=prediction;
band x=fixation upper=ucl lower=lcl /
        fill transparency=.5
        legendlabel="individuelle";
band x=fixation upper=uclm lower=lclm /
        fill transparency=.1
        legendlabel="moyenne";
series x=fixation y=predicted /
legendlabel="prédiction";
yaxis label="intention d'achat";
xaxis label="temps de fixation (en secondes)";
run;
```

Intervalles de confiance pour la moyenne et intervalles de prédiction



L'intervalle de confiance pour la moyenne (gris foncé) est plus étroit que l'intervalle de prédiction (gris clair). Les intervalles de prédiction et de confiance deviennent plus larges à mesure que l'on s'éloigne de la valeur moyenne de `fixation`.