Modélisation - Projet partie 1

Hilaire Touyem Xavier Lapointe Georges Balogog
Xavier Péladeau

2023-10-13

Table of contents

Analyse exploratoire

Distribution des durées des déplacements

Nous procédons d'abord à une analyse exploratoire des données pour mieux comprendre la distribution des variables et les relations entre elles.

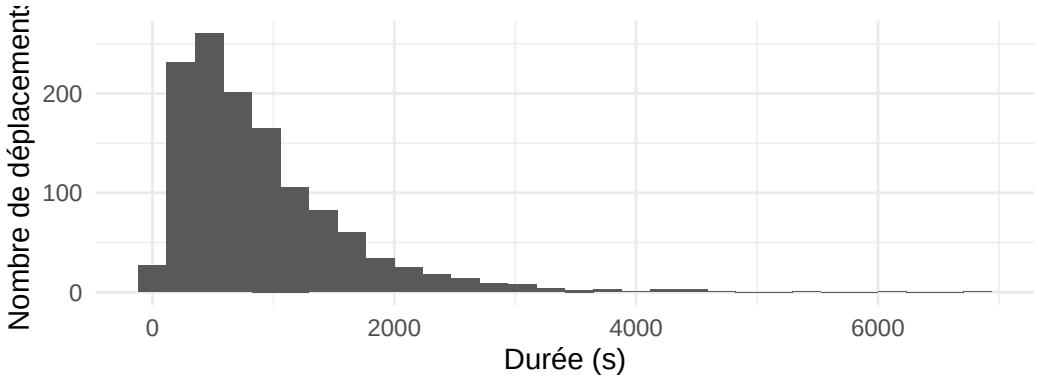

Figure 1: Distribution des durées des déplacements.

La Figure 1 affiche la distribution de la variable-réponse (durées des déplacements en seconde) observées dans l'échantillon. Celle-ci semble être asymétrique à droite. Comme les modèles de régression linéaires que nous utiliserons supposent une distribution normale des résidus, il pourrait être intéressant de considérer une transformation logarithmique pour la variable-réponse.

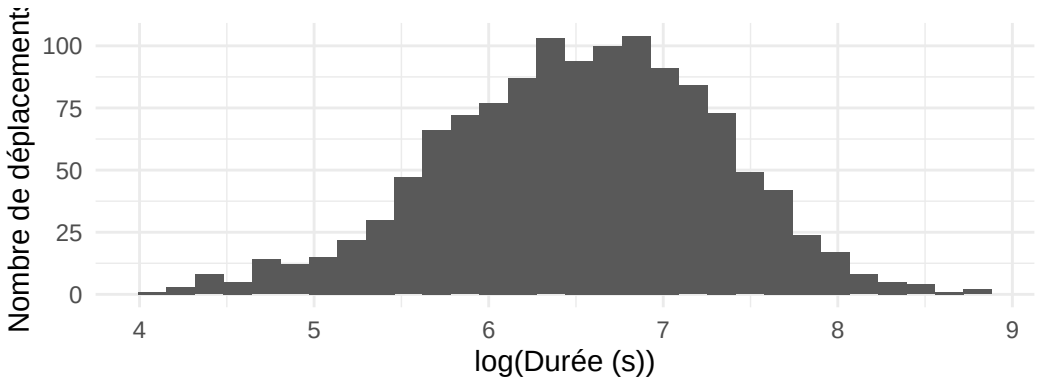

Figure 2: Distribution du logarithme des durées des déplacements.

La Figure 2 affiche la distribution de la variable-réponse après transformation logarithmique. Celle-ci semble plus symétrique et est être plus appropriée pour les modèles de régression

linéaire. Nous utiliserons donc le logarithme des durées des déplacements comme variable-réponse dans nos modèles.

Résumé des variables explicatives

Variables catégorielles

La Figure 3 présente le nombre d'observations pour chaque sous-groupe des variables catégorielles du jeu de données, afin de vérifier que les catégories sont bien balancées. Le jeu de données contient trois variables catégorielles, soient `jour`, `mem`, et `pointe`.

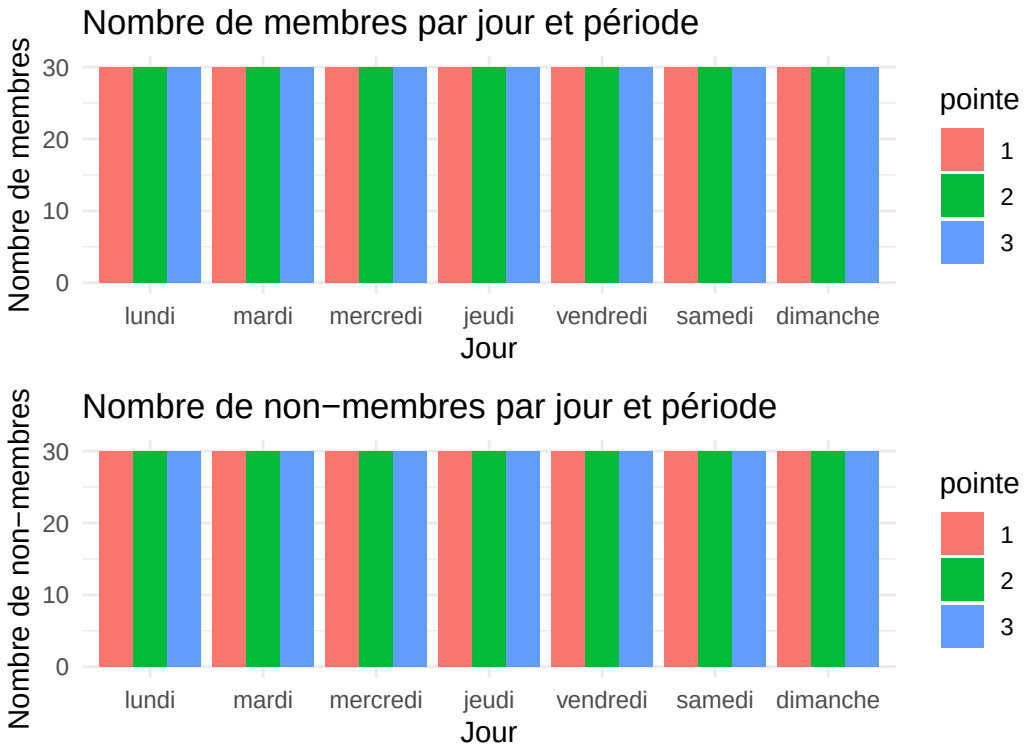

Figure 3: Distribution des variables catégorielles du jeu de données.

Variables continues

Les graphiques de la Figure 4 présente les statistiques descriptives des variables explicatives continues du jeu de données, soit `temp`, `prec` et `dep`. Ces statistiques permettent de mieux comprendre la distribution des variables continues et de détecter des valeurs aberrantes.

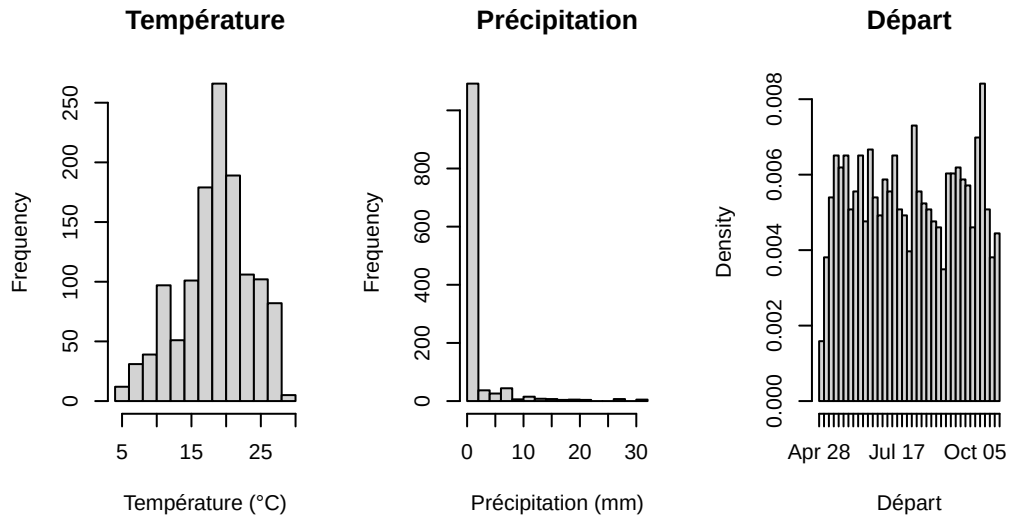

Figure 4: Distribution des variables explicatives continues du jeu de données.

Question 1.1 : En moyenne, les membres de BIXI effectuent-ils des trajets plus courts que les non-membres?

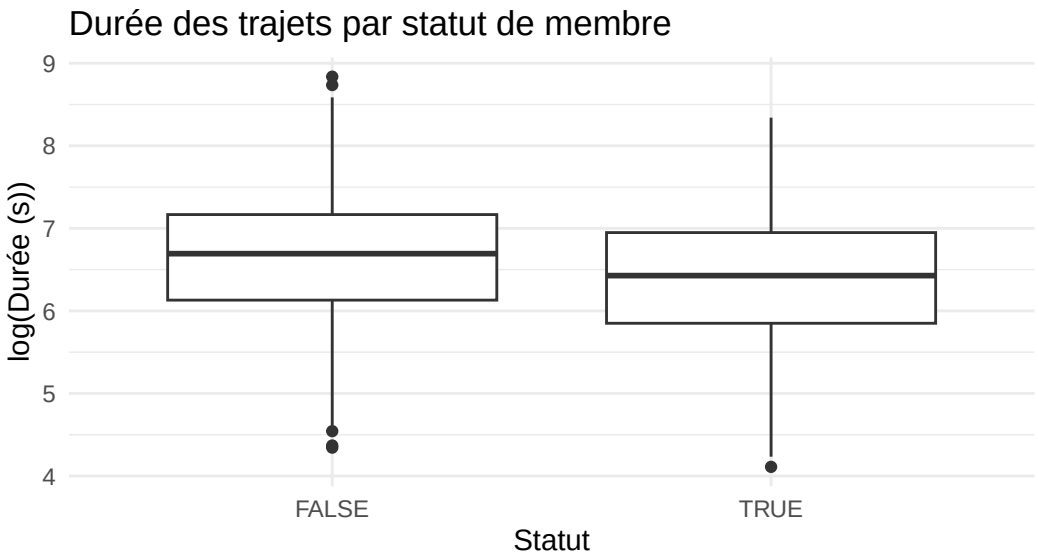

Figure 5: Relation entre le statut de membre et la durée des trajets.

La Figure 5 illustre la relation entre le statut de membre et la durée des trajets. Les boîtes suggèrent que les membres de BIXI effectuent des trajets légèrement plus longs que les non-

membres. Cependant, pour confirmer cette tendance, nous devons effectuer une analyse statistique.

Pour déterminer si les membres BIXI effectuent des trajets plus courts que les non-membres, nous définissons un modèle de régression linéaire modélisant uniquement la relation entre le statut de membre et la durée du trajet (Modèle 1) : $log(dur_i) = \beta_0 + \beta_1 X_i + \epsilon_i$

Où:

- $log(dur_i)$ représente la log-durée d'un trajet spécifique.
- X indique si la personne est membre (1) ou non-membre (0) de BIXI.
- β_0 est la durée moyenne des trajets pour les non-membres.
- β_1 montre comment le fait d'être membre change la log-durée du trajet.
- ϵ_i est le terme d'erreur suivant une distribution normale $N(0, \sigma^2)$.

Notre hypothèse nulle est qu'il n'y a pas de différence significative dans la durée moyenne des trajets entre les membres et les non-membres de BIXI: $\mathbb{H}_0 : \beta_1 = 0$ $\mathbb{H}_1 : \beta_1 \neq 0$

```
model <- lm(log_dur ~ mem, data = data)
coefs <- summary(model)$coefficients
intervals <- confint(model)
```

Après avoir ajusté le modèle à notre échantillon, les paramètres prennent les valeurs suivantes:

$\hat{\beta_0} = 6.6495$ $\hat{\beta_1} =$ -0.2541

La valeur négative prise par $\hat{\beta_1}$ (IC 95% [-0.3395, -0.1687]) indique que les usagers qui sont membres effectuent des trajets plus courts que les non-membres. Précisément, exp(-0.2541) - 1 78%, signifiant que les membres font des trajets 22.4% plus courts en moyenne. En tenant compte de l'intervalle de confiance, cette réduction de la durée des trajets pour les membres se situe entre 15.5% et 28.8% avec 95% de confiance.

En conclusion, l'analyse statistique nous permet de rejeter l'hypothèse nulle avec un haut degré de confiance ($p < 0.001$). Nous pouvons donc conclure que oui, il existe une différence significative dans la durée moyenne des trajets entre les membres et les non-membres de BIXI.

Question 1.2 : Les résultats sont-ils les mêmes si l'on tient compte de l'utilisation en fin de semaine ou en semaine?

Pour déterminer s'il y a une différence de durée de trajet en fonction du statut de membre et du moment de la semaine, nous définissons le modèle de régression linéaire suivant (**Modèle 2**) :

$$log(\text{dur}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + \epsilon_i$$

Où :

- dur_i est la durée du trajet i.
- X_1 indique si la personne est membre (1) ou non-membre (0) de BIXI.
- X_2 indique si le trajet a lieu le weekend (1) ou en semaine (0).
- β_0 est la durée moyenne des trajets pour les non-membres en semaine.
- β_1 montre comment le fait d'être membre change la durée du trajet.
- β_2 montre comment le fait d'être le weekend change la durée du trajet.
- β_3 montre si l'effet d'être membre est différent le weekend par rapport à la semaine.
- ϵ_i est le terme d'erreur suivant une distribution normale $N(0, \sigma^2)$.

```
model <- lm(log_dur ~ mem * is_weekend, data = data)
coefs <- summary(model)$coefficients
intervals <- confint(model)
```

Après avoir ajusté le modèle à notre échantillon, les paramètres prennent les valeurs suivantes:

$\hat{\beta_0} = 6.6148$ $\hat{\beta_1} = -0.2293$ $\hat{\beta_2} = 0.1215$ $\hat{\beta_3} = -0.087$

En semaine, les membres font des trajets environ 20.5% (Calcul : exp(-0.22926) - 1) * 100). plus courts que les non-membres (coefficient -0.22926, IC 95% [-0.33014, -0.12838], $p < 0.001$). Cela correspond à une réduction de la durée du trajet entre 11.9% et 28.1% avec 95% de confiance.

Le weekend, les membres font des trajets environ 17.7% plus courts que les non-membres (Calcul : exp(-0.22926 + 0.12153 - 0.08702) - 1) * 100). L'effet du weekend (coefficient 0.12153, IC 95% [-0.01193, 0.25499], $p = 0.0745$) n'est pas statistiquement significatif au seuil de 0.05. De même, l'interaction entre le statut de membre et le weekend (coefficient -0.08702, IC 95% [-0.27577, 0.10173], $p = 0.3664$) n'est pas statistiquement significative. En conclusion, nous ne

pouvons pas rejeter l'hypothèse nulle principale ou secondaire. Bien que nous observions une tendance générale où les membres effectuent des trajets significativement plus courts, nous ne pouvons pas affirmer avec certitude que le moment de la semaine ou son interaction avec le statut de membre influencent significativement la durée des trajets.

Question 2.1 : Est-ce que la durée des trajets est influencée par la météo?

Deux variables météorologiques sont inclusent dans le jeu de données, soient la température (`temp`) et les précipitations (`prec`). Nous effectuerons un test pour chacune d'entre elles afin de déterminer si elles influencent la durée moyenne des trajets.

Impact de la température

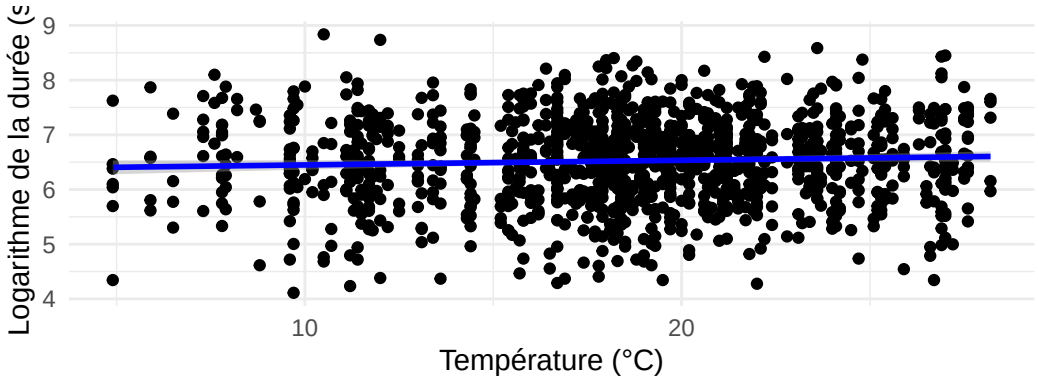

Figure 6: Relation entre la température et la durée des trajets.

La Figure 6 illustre la relation entre la température et la durée des trajets. La droite de régression linéaire suggère une légère augmentation de la durée des trajets avec la température.

Pour tester l'impact de la température sur la durée des trajets, nous définissons le modèle de régression linéaire suivant (**Modèle 3**) : $log(\text{dur}_i) = \beta_0 + \beta_1 \text{temp}_i + \epsilon_i$

Notre hypothèse nulle est que la température n'a pas d'effet significatif sur la durée moyenne des trajets Bixi :

$\mathbb{H}_0 : \beta_1 = 0$

```
model <- lm(log_dur ~ temp, data = data)
coefs <- summary(model)$coefficients
intervals <- confint(model)
```

L'ordonnée à l'origine (β_0) de 6,363649 indique que lorsque la température est à 0°C, la durée moyenne prédite du trajet est d'environ 580,3 secondes (Calcul : exp(6,363649)). Le coefficient de température (β_1) de 0,008534 révèle que pour chaque augmentation d'un degré Celsius, la durée du trajet augmente en moyenne de 0,857% (Calcul : (exp(0,008534) - 1) * 100). Ce coefficient n'est pas significatif (p-value = 0,0535), avec un intervalle de confiance à 95% allant de -0,014% à 1,736% d'augmentation.

Impact des précipitations

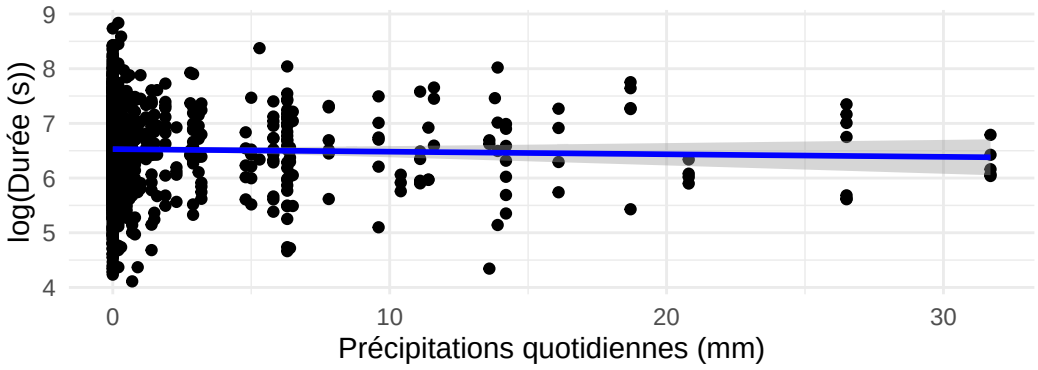

Figure 7: Relation entre la température et la durée des trajets.

La Figure 7 illustre la relation entre les précipitations et la durée des trajets. La droite de régression linéaire suggère une très légère diminution de la durée des trajets avec les précipitations.

Pour tester l'impact des précipitations sur la durée des trajets, nous définissons le modèle de régression linéaire suivant (**Modèle 4**) : $log(\text{dur}_i) = \beta_0 + \beta_1 \text{prec}_i + \epsilon_i$

Notre hypothèse nulle est que le niveau de précipitation n'a pas d'effet significatif sur la durée moyenne des trajets Bixi :

$\mathbb{H}_0 : \beta_1 = 0$ $\mathbb{H}_1 : \beta_1 \neq 0$

```
model <- lm(log_dur ~ prec, data = data)
```

L'ordonnée à l'origine (β_0) de 6,528871 indique que lorsque les précipitations sont à leur valeur de référence dans le modèle, la durée moyenne prédite du trajet est d'environ 685,1 secondes (Calcul : exp(6,528871)). Quant au coefficient de précipitations (β_1) de -0,004693, il suggère une diminution moyenne de 0,468% de la durée du trajet pour chaque millimètre de précipitations supplémentaire (Calcul : (exp(-0,004693) - 1) * 100). Cependant, ce coefficient n'est pas statistiquement significatif (p-value = 0,389), avec un intervalle de confiance à 95% allant de -1,546% à 0,617%.

Question 2.2 : Au vu du résultat que vous obtenez, est-ce que vos modèles initiaux devraient être revisités?

Étant donné que les variables température et précipitations se sont révélées statistiquement non significatives dans leurs modèles respectifs, il n'est pas justifié de les incorporer au modèle initial. Cette conclusion s'applique également à la variable binaire distinguant les jours de semaine des week-ends, qui s'est aussi avérée non significative dans l'analyse précédente.

Question 3.1 : Les durées de trajets sont-elles différentes selon que l'on se trouve aux heures de pointe ou non en semaine?

Comme cette question traite seulement des tendances lors des jours de semaine, nous excluons d'abord les observations des fins de semaine :

```
data_week <- data %>% filter(!is_weekend)
```

Nous conservons le statut du membre comme variable explicative, comme les membres semblent avoir des comportements significativement différents des non-membres, à un seuil de confiance de 0.05. Nous sommes intéressés à comparer la durée moyenne des trajets de notre échantillon effectués pendant les heures de pointe à la durée moyenne hors des heures de pointe. À cette fin, nous considérons les hypothèses suivantes:

$\mathbb{H}_0 : \mu_{\text{pointe\_membre}} + \mu_{\text{pointe\_non-membre}} = \mu_{\text{non-pointe\_membre}} + \mu_{\text{non-pointe\_non-membre}}$

$\mathbb{H}_1 : \mu_{\text{pointe\_membre}} + \mu_{\text{pointe\_non-membre}} \neq \mu_{\text{non-pointe\_membre}} + \mu_{\text{non-pointe\_non-membre}}$

Où: μ_x = indique la durée moyenne des trajets pour le sous-groupe x.

Pour tester ces hypothèses, nous ajustons le modèle suivant (**Modèle 4**) : $\mathbb{E}[\ log(dur) \mid .] = \beta_0 + \beta_1.\mathbb{1}_{\text{IsMem}} + \beta_2.\mathbb{1}_{\text{Pointe = 2 = Soir}} + \beta_3.\mathbb{1}_{\text{Pointe = 3 = Hors pointe}}$

```
contrast                              estimate    SE  df t.ratio p.value
Pointe Semaine VS. NotPointeSemaine     -0.269 0.109 896  -2.467  0.0138
```

Heures pointe en semaine (1+2+4+5) vs. **Pas de pointe en semaine** (3+6) " ; ceci correspond a l'egalite des moyennes telle que

$(1.\mu * 000 + 1. * \mu 010 + 1.\mu * 100 + 1. * \mu 110) = (2.\mu * 001 + 2. * \mu 101)$

en ordant cela (R fonctionne par ordre alpha-numerique), on ecrit:

$1.\mu * 000 + 1. * \mu 010 - 2.\mu * 001 + 1. * \mu 100 + 1.\mu * 110 - 2. * \mu 101 = 0 \quad [c_1 = (1, 1, -2, 1, 1, -2)]$

On ecrit donc le vecteur de contraste : 'c_1 = (1,1,-2,1,1,-2)' sachant que le modele est ordonne par : '000' (groupe de reference) + '010' + '001' + '100' + '110' + '101'.

L'analyse révèle des différences significatives dans la durée des trajets selon les périodes de pointe et l'appartenance : Estimation : -0.2694, p-value : 0.0138 (significatif). Les trajets en période de pointe sont plus courts que ceux hors pointe, suggérant un comportement différent des utilisateurs.

Question 3.2 : Existe-t-il des différences entre l'utilisation pour les heures de pointes en semaine le matin ou le soir?

```
 contrast                       estimate    SE  df t.ratio p.value
 Pointe Matin   vs. PointeSoir   0.0486 0.063 896   0.771  0.4407
```

Nous sommes intéressés à comparer la durée moyenne des trajets de notre échantillon effectués pendant les heures de pointe du matin à la durée moyenne lors des heures de pointe du soir. À cette fin, nous considérons les hypothèses suivantes:

$\mathbb{H}_0 : \mu_{\text{matin\_membre}} + \mu_{\text{matin\_non-membre}} = \mu_{\text{soir\_membre}} + \mu_{\text{soir\_non-membre}}$

$\mathbb{H}_1 : \mu_{\text{matin\_membre}} + \mu_{\text{matin\_non-membre}} \neq \mu_{\text{soir\_membre}} + \mu_{\text{soir\_non-membre}}$

Où: $\mu_x =$ indique la durée moyenne des trajets pour le sous-groupe x.

Estimation : 0.0486, p-value : 0.4407 (non significatif). Aucune différence notable entre les trajets matin et soir durant les périodes de pointe.

On veut aussi tester, " **Heures pointe Matin semaine** (1+4) vs. **Heures pointe Soir semaine** (2+5)" ; ceci correspond a l'egalite des moyennes telle que

$(1.\mu 000 + 1.\mu 100) = (1.\mu 010 + 1.\mu 110) + 0.(\mu 001 + \mu 101)$

en ordant cela (R fonctionne par ordre alpha-numerique), on ecrit:

$1.\mu_{000} - 1.\mu_{010} - 0.\mu_{001} + 1.\mu_{100} - 1.\mu_{110} - 0.\mu_{101} = 0 \quad [c_2 = (1, -1, 0, 1, -1, 0)]$

On ecrit donc le vecteur de contraste : 'c_2 = (1,-1,0,1,-1,0)' sachant que le modele est ordonne par : '000' (groupe de reference) + '010' + '001' + '100' + '110' + '101'.

Effets des Moyennes Estimées Les moyennes pour les membres et non-membres varient légèrement, avec les non-membres ayant tendance à des durées de trajet plus longues dans certaines catégories de pointe.

Conclusion Les résultats soulignent une dynamique entre appartenance et période de pointe, avec des implications pour la gestion des trajets et l'engagement des utilisateurs.s

Voici l'interpretation des coefficients :

$$
\begin{cases}
\mu_{000} = \beta_0 & & & & \text{(1) Pas Membre + Matin} \\
\mu_{010} = \beta_0 & & + \beta_2 & & \text{(2) Pas Membre + Soir} \\
\mu_{001} = \beta_0 & & & + \beta_3 & \text{(3) Pas Membre + Hors pointe} \\
\mu_{100} = \beta_0 + \beta_1 & & & & \text{(4) Membre + Matin} \\
\mu_{110} = \beta_0 + \beta_1 & + \beta_2 & & & \text{(5) Membre + Soir} \\
\mu_{101} = \beta_0 + \beta_1 & & & + \beta_3 & \text{(6) Membre + Hors pointe}
\end{cases}
$$