Instructions:

- Répondez aux questions suivantes à l'aide de SAS et fournissez le code utilisé pour vos analyses dans un fichier (extension `.txt`, encodage `utf8`).
- Votre rapport doit être remis en ligne en format PDF et en version papier en classe et ne devrait pas faire plus de 15 pages; toute page excédentaire sera ignorée lors de la correction. Soyez concis mais précis: n'incluez que les sorties pertinentes.
- Écrivez les hypothèses en fonction des paramètres du modèle, rapportez la statistique de test pertinente, sa loi nulle et la conclusion du test dans le contexte du problème.
- Les erreurs sont pénalisées même si elles ne sont pas en lien direct avec la question.
- Les (sous)-questions avec une étoile ($\star$) sont à compléter en équipe de deux ou trois. Inclure avec le nom de votre coéquipier/coéquipière, les graphiques et l'analyse dans un seul des deux rapports)

1.1 **Graphiques** ($\star$). Trouvez deux graphiques en ligne sur un journal et publiez le lien vers votre graphique/article sur Piazza. Vous ne pouvez pas utiliser l'un des exemple couvert en classe ou déjà proposé par un.e autre étudiant.e du cours (premier arrivé, premier servi).[1] Des points seront accordés pour l'originalité.
Discutez brièvement des éléments suivants:

- Résumez en vos propres mots l'histoire racontée par le graphique
- Quel type de graphique a été employé (géométrie): est-ce que ce choix est approprié dans le contexte?
- Identiez les types de variables et les applications (axe des x, axe des y, forme, couleur, etc.)
- Identifiez les points forts et faibles du graphique.

1.2 **Calcul de la puissance d'un test statistique**
Le programme SAS `puissance.sas`, écrit par hrefhttps://blogs.sas.com/content/iml/2020/08/12/simulation-estimate-power-of-test.htmlRick Wicklin de SAS Institute Inc., contient du code pour faire une étude de simulation afin de calculer la puissance d'un test-t pour deux échantillons (l'équivalent d'un modèle linéaire simple avec une variable binaire comme variable explicative).

(a) Expliquez brièvement dans vos mots en quoi consiste les étapes de l'étude de simulation.
(b) Tracez le graphique pour les paramètres $n_1 = n_2 = 10$, $\sigma = 1$ et $B = 10000$ et commentez sur l'apparence de ce dernier.
(c) Faites varier le nombre de simulations de $B = 100$ à $B = 10000$. Que remarquez-vous quand le nombre de simulation est petit? Expliquez pourquoi cet effet disparaît quand la nombre de simulations augmente.
(d) Modifiez le code pour que la taille des groupes soit $n_1 = 10, n_2 = 30$ et $n_1 = 20, n_2 = 20$. Dans lequel des deux scénarios la puissance est-elle plus élevée et pourquoi?
(e) Modifiez le code pour simuler $n = 20$ observations dans chaque groupe de lois normales d'écart-type $\sigma_1 = 1$ et $\sigma_2 = 5$. Quel est l'estimé du niveau de votre test? Rapportez l'estimé avec le nombre de simulations, un intervalle de confiance ponctuel à 90% et expliquez comment vous avez dérivé ce dernier.

1.3 **Modèle linéaire pour les données `diamants`** Le jeu de données `diamants` contient le prix de 1000 diamants et les variables suivantes:

- `prix`: prix du diamant (en dollars US)
- `carat`: masse du diamant (en carats)
- `taille`: qualité de la taille (`assez bonne`, `bonne`, `tres bonne`, `excellente`, `ideale`)
- `couleur`: couleur du diamant, soit `transparent` (couleurs DEF) ou `presque transparent` (GHIJ)
- `clarte`: clarté du diamant, un parmi `I1` (pire), `SI2`, `SI1`, `VS2`, `VS1`, `VVS2`, `VVS1`, `IF` (meilleure)
- `longueur`: longueur du diamant (en mm)
- `largeur`: largeur du diamant (en mm)
- `profondeur`: profondeur du diamant (en mm)

[1] Quelques suggestions: les fils Twitter de la BBC, du Washington Post, du New York Times, etc., les articles de magazines ou des sources officielles comme Statistique Canada ou le US Census Bureau. Évitez aussi les séries chronologiques génériques (par ex., le cours d'une action à la bourse).

Compilé le 06/10/2020 à 22:51

- `profondeurtot`: profondeur totale du diamant, soit 2`profondeur`/(`longueur` + `largeur`) (en pourcentage)
- `table`: largeur du dessus du diamant par rapport à son diamètre

(a) (⋆) Faites une analyse exploratoire des données (une page maximum de texte). Résumez les éléments clés nécessaires à une bonne interprétation des données. En particulier, considérez les points suivants
- Y a-t-il des valeurs aberrantes ou des erreurs de prétraitement dans la base de donnée?
- Quelle est la relation entre carat et prix?
- Quelles variables explicatives sont fortement corrélées entre elles?
- On aurait pu ajouter les dimensions comme variables explicatives au modèle comme variables continues: est-ce que ce serait logique ou pas?

(b) Ajustez un modèle linéaire pour le prix du diamant avec `carat`, `taille`, `couleur` et `clarte`. Utilisez `presque transparent`, `I1`, `assez bonne` comme catégories de référence.
i. Interprétez les paramètres pour `taille` (`excellent`) et `carat`.
ii. Interprétez le coefficient de la moyenne associé à `couleur` (`transparent`). Est-ce que la valeur de l'estimé vous semble logique? Expliquez.
iii. Est-ce que l'augmentation du prix est la même pour chaque échelon de clarté?
iv. Produisez des diagnostics graphiques des résidus et commentez sur la validité des postulats du modèle linéaire.

(c) On choisit de modéliser plutôt ln(`prix`) avec les même variables explicatives.
i. Est-il nécessaire de transformer les variables explicatives? Si oui, et procédez à la transformation si nécessaire et justifiez votre choix. Sinon, expliquez votre raisonnement.
ii. Interprétez les paramètres pour `taille` (`très bonne`) et `carat` (à l'échelle des données en USD).
iii. Quel modèle vous semble le plus adéquat? Justifiez votre réponse à l'aide de diagnostics graphiques et de tests (ou d'autres critères).

1.4 Les données `renfe` contiennent des informations sur 10 000 billets de trains vendus par la compagnie Renfe, l'entreprise ferroviaire publique espagnole. Les données incluent les variables:
- `prix`: prix du billet (en euros);
- `tarif`: variable catégorielle indiquant le tarif du billet, un parmi `AdultoIda`, `Promo` et `Flexible`;
- `classe`: classe du billet, soit `Preferente`, `Turista`, `TuristaPlus` ou `TuristaSolo`;
- `type`: variable catégorielle indiquant le type de train, soit Alta Velocidad Española (`AVE`), soit Alta Velocidad Española conjointement avec TGV (un partenariat entre la SNCF et Renfe pour les trains à destination ou en provenance de Toulouse) `AVE-TGV`, soit les trains régionaux `REXPRESS`; seuls les trains étiquetés `AVE` ou `AVE-TGV` sont des trains à grande vitesse.
- `jour` entier indiquant le jour de la semaine du départ allant de dimanche (`1`) à samedi (`7`).

Ajustez un modèle linéaire pour expliquer le `prix` des billets de trains à grande vitesse en fonction de `classe`, `tarif`, d'une interaction entre les deux, et d'une variable additionnelle indiquant si le jour de départ est une fin de semaine ou pas (avec commes catégories de référence `Flexible`, `Preferente`, et jour de semaine).

(a) Écrivez l'équation du modèle théorique postulé.
(b) Est-ce que le terme d'interaction entre `classe` et `tarif` est statistiquement significatif?
(c) Rapportez les estimés des paramètres de la moyenne et interprétez les coefficients associés aux variables explicatives `classe`, `tarif` (incluant les termes d'interaction entre les deux).
(d) Expliquez pourquoi on ne considère pas le test F pour `classe` dans le modèle avec interaction: que représente-t-il dans le contexte?
(e) Testez la significativité globale du modèle.
(f) Prédisez le prix d'un billet `AVE` au tarif `Promo` et `Turista` circulant un samedi et fournissez un intervalle de prédiction à 90%.
(g) Produisez des diagnostics graphiques des résidus et commentez sur la validité des postulats du modèle linéaire.

Compilé le 06/10/2020 à 22:51

1.5 On considère un modèle pour le nombre de ventes quotidiennes dans un magasin, supposées indépendantes. Votre gestionnaire vous indique que ce dernier fluctue selon qu'il y ait des soldes ou pas. Vous spécifiez un modèle de Poisson avec fonction de masse

$$\mathsf{P}\left(Y_i = y_i \mid x_i\right) = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, \ldots$$

où $\lambda_i = \exp(\beta_0 + \beta_1 \texttt{solde}_i)$ et $\texttt{solde}_i$ est un indicateur binaire qui vaut un lors des soldes et zéro autrement.

(a) Dérivez l'estimateur du maximum de vraisemblance pour (β_0, β_1).

(b) Calculez les estimés si on a un échantillon de 12 observations, soit le nombre de ventes hors solde, $\{2; 5; 9; 3; 6; 7; 11\}$, et pendant les soldes, $\{12; 9; 10; 9; 7\}$.

(c) Calculez la matrice d'information observée et utilisez-la pour dériver les erreurs-type de $\widehat{\beta}_0$ et de $\widehat{\beta}_1$ et un intervalle de confiance à niveau 95%.

(d) Votre gérant veut savoir si le profits quotidiens moyens pendant les soldes sont différentes des jours ordinaires, sachant que le profit moyen pour les jours hors solde est de 25$ par transaction, mais seulement 20$ pendant les soldes. Faites un test de rapport de vraisemblance (profilée) pour estimer cette hypothèse.

Compilé le 06/10/2020 à 22:51