TD 3 : Sélection de variables

Exercice 1. Fisher global

1. Rappeler la définition des lois du chi-deux, de Student, et de Fisher.
2. On se place dans le cadre d'une régression linéaire gaussienne multiple $Y = X\beta + \varepsilon$. Rappeler les expressions et les lois de $\hat{\beta}$ et $\hat{\sigma}$.
3. En déduire que la statistique de test de Fisher global
$$F = \frac{1}{p\hat{\sigma}^2}\hat{\beta}^T(X^TX)\hat{\beta}$$
suit sous H_0 une loi de Fisher de paramètre p, $n-p$.
On rappelle que si $N \sim \mathcal{N}(\mu, \Sigma)$ est un vecteur gaussien dans $\mathbb{R}^k$ avec Σ inversible, alors $(N-\mu)^T\Sigma^{-1}(N-\mu)$ suit une loi du chi-deux à k degrés de liberté.
4. Réécrire F en fonction de Y et $\hat{Y}$.
5. Que teste cette statistique ? Que peut-on en dire en pratique ?

Exercice 2. Fisher emboîté

1. Rappeler le test entre modèles emboîtés et donner la statistique de test F en fonction de Y, $\hat{Y}$, et $\hat{Y}_0$. Dans quel contexte retrouve-t-on l'expression de l'exercice 1 ?
2. Réécrire cette quantité en fonction de SCR et SCR_0.
3. Montrer que
$$F = \frac{n-p}{q}\frac{R^2 - R_0^2}{1-R^2}\,,$$
où R^2 et R_0^2 sont les coefficients de détermination associés respectivement au modèle complet et au modèle emboîté.

Exercice 3. Équivalence des tests de Student et de Fisher

On souhaite montrer l'équivalence entre les tests de Student et de Fisher pour la nullité d'un paramètre. On considère le modèle
$$Y = X\beta + \varepsilon\,,$$
et on souhaite tester la nullité du dernier coefficient β_p.

1. Rappeler la statistique de test T du test de Student sous l'hypothèse $H_0 : \beta_p = 0$.
2. Rappeler la statistique de test F du test de Fisher pour les modèles emboîtés.
3. Montrer que si T_n est une variable aléatoire de loi de Student à n degrés de liberté, alors T_n^2 suit une loi de Fisher. Avec quels degrés de liberté ?
4. En notant la matrice X par blocs sous la forme $X = [X_0|X_p]$, avec $X_0 = [X_1|\cdots|X_{p-1}]$ les $p-1$ premières colonnes de X, donner sous forme de blocs l'expression de la matrice X^TX.
5. Soit M une matrice inversible écrite sous forme de blocs :
$$M = \left(\begin{array}{c|c} T & U \\ \hline V & W \end{array}\right)$$
avec T inversible. Alors $Q = W - VT^{-1}U$ est inversible et l'inverse de M est
$$M^{-1} = \left(\begin{array}{c|c} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ \hline -Q^{-1}VT^{-1} & Q^{-1} \end{array}\right)$$

À l'aide de ce résultat, montrer que

$$[(X^T X)^{-1}]_{pp} = (X_p^T (I_n - P_0) X_p)^{-1} ,$$

où P_0 est la matrice $n \times n$ de projection orthogonale sur l'espace $\mathcal{M}_0$ engendré par les colonnes de X_0.

6. On note $\hat{Y}$ (resp. $\hat{Y}_0$) le projeté orthogonal de Y sur $\mathcal{M}$ (resp. $\mathcal{M}_0$). Montrer que

$$\hat{Y} - \hat{Y}_0 = \hat{\beta}_p (I_n - P_0) X_p .$$

7. En déduire que $F = T^2$ et conclure.

Exercice 4. Estimation sous contrainte Dans le modèle de régression linéaire, il arrive parfois que l'on souhaite imposer des contraintes linéaires à β, par exemple que sa première coordonnée soit égale à 1. Nous supposerons en général que nous imposons q contraintes linéairement indépendantes à β, ce qui s'écrit sous la forme : $R\beta = r$, où R est une matrice $q \times p$ de rang $q < p$ et r un vecteur de taille q. Montrer que l'estimateur des moindres carrés sous contraintes s'écrit:

$$\hat{\beta}_c = \hat{\beta} + (X^T X)^{-1} R^T \Big[R (X^T X)^{-1} R^T \Big]^{-1} (r - R\hat{\beta}) .$$

Exercice 5. Modèle de Cobb-Douglas Nous disposons pour n entreprises de la valeur du capital K_i, de l'emploi L_i, et de la valeur ajoutée V_i. Nous supposons que la fonction de production de ces entreprises est du type Cobb-Douglas :

$$V_i = \lambda L_i^{\beta} K_i^{\gamma} .$$

1. Comment se ramène-t-on à un modèle de régression linéaire ?
2. Pour $n = 1658$ entreprises, nous avons obtenu les estimateurs suivants :

$$\hat{\beta} = \begin{pmatrix} 3.136 \\ 0.738 \\ 0.282 \end{pmatrix} ,$$

avec $R^2 = 0.945$ et $SCR = 148.27$. Nous donnons aussi

$$(X^T X)^{-1} = \begin{pmatrix} 0.0288 & 0.0012 & -0.0034 \\ 0.0012 & 0.0016 & -0.0010 \\ -0.0034 & -0.0010 & 0.0009 \end{pmatrix} \text{ et } X^T X = \begin{pmatrix} 423 & 2231 & 4077 \\ 2231 & 13808 & 23769 \\ 4077 & 23769 & 42923 \end{pmatrix} .$$

Calculer $\hat{\sigma}^2$ et une estimation de la variance de $\hat{\beta}$.

3. Donner un intervalle de confiance au niveau 95% pour β. Même question pour γ.
4. Tester au niveau 5% l'hypothèse $H_0 : \gamma = 0$ contre $H_1 : \gamma > 0$.
5. Nous voulons tester l'hypothèse selon laquelle les rendements d'échelle sont constants (une fonction de production F est à rendement d'échelle constant si pour tout $\beta \geq 0$,

$$F(\theta L, \theta K) = \theta F(L, K) .$$

Quelles sont les contraintes vérifiées par le modèle lorsque les rendements d'échelle sont constants ? Tester au niveau 5%

$$H_0 : \text{ les rendements sont constants} \quad \text{vs} \quad H_1 : \text{ les rendements sont croissants.}$$