



NOM :

PRÉNOM :

Questions de cours

1. On a

$$X'X = n \begin{pmatrix} 1 & 0 \\ 0 & \text{Var}(X) \end{pmatrix} \text{ et } \text{Var}(\hat{\theta}) = \sigma^2(X'X)^{-1}.$$

Déduire la relation entre la matrice de variance-covariance de $\hat{\theta}$ et celle de X . Quelle est la condition nécessaire et suffisante pour que les estimateurs des coefficients des variables ne covarient pas?

2. Imaginons qu'un modèle soit globalement significatif. Est-il possible que les tests de significativité partielle se révèlent tous négatifs?

Exercice 1. On dispose d'un jeu de données (X, Y, Z) suivant.

$$X = (1, 2, 3, 4, 5) \quad Y = (9, 13, 2, 8, 1) \quad Z = (3, 4, 5, 6, 8)$$

Étape 1

Dans un premier temps, nous considérons le modèle à deux variables explicatives. Les résultats sont présentés dans les tableaux suivants.

Call:

`lm(formula = z ~ x + y)`

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.91629	0.85672	2.237	0.1548
x	1.14851	0.18001	6.380	0.0237 *
y	-0.02452	0.05659	-0.433	0.7071

Analysis of Variance Table

Response: z

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	14.4000	14.4000	78.7586	0.01246 *
y	1	0.0343	0.0343	0.1877	0.70706
Residuals	2	0.3657	0.1828		

1. Réécrivez le modèle en utilisant les résultats de régression.
2. Calculez le coefficient de détermination. Ce modèle est-il statistiquement satisfaisant?
3. Une variable ne doit elle pas être éliminée? Laquelle?

Étape 2

Dans la seconde étape, nous considérons le modèle à une seule variable explicative. Les résultats sont présentés dans les tableaux suivants.

Call:

```
lm(formula = z ~ x)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.6000	0.3830	4.178	0.0250 *
x	1.2000	0.1155	10.392	0.0019 **

Analysis of Variance Table

Response: z

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	14.4	14.4000	108	0.001901 **
Residuals	3	0.4	0.1333		

Call:

```
lm(formula = z ~ y)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.9348	1.2810	5.414	0.0124 *
y	-0.2628	0.1604	-1.639	0.1997

Analysis of Variance Table

Response: z

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
y	1	6.9917	6.9917	2.6863	0.1997
Residuals	3	7.8083	2.6028		

4. Vérifiez que le coefficient de détermination partielle associé à X est beaucoup plus élevé que celui de Y .

5. Testez la position du coefficient de X par rapport à 0.8 avec le niveau de test $\alpha = 10\%$.

6. Réécrivez le modèle $Z = f(X)$ en utilisant les résultats de régression. Calculez son coefficient de détermination.

Étape 3

Dans la dernière étape, nous considérons le modèle log-linéaire ci dessous.

$$\ln Z = \mu + \beta X + \varepsilon.$$

Nous avons les résultats suivants.

$$\ln Z = (1.1, 1.4, 1.6, 1.8, 2.1), \quad \varepsilon = (-0.021, 0.030, 0.016, -0.038, 0.013)$$

7. Le modèle log-linéaire est-il le meilleur que le modèle linéaire simple dans l'étape 2?

Exercice 2. Un statisticien a effectué l'analyse du résidu sur quatre jeux de données. Les figures ci dessous représentent les nuages de points des résidus calculés par rapport aux variables prévues. Elles vous paraissent-elles indiquer quelques anomalies?

