Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2015/2016

Sujet 14: Test sur le coefficient directeur d'une droite de régression

---

• **Quelques rappels sur la définition mathématique des tests d'hypothèses**:
Soit X une variable aléatoire à valeurs dans un ensemble $\mathcal{X}$, de loi P_θ où $\theta \in \Theta$. Supposons que l'on veuille effectuer un test statistique d'hypothèses portant sur θ. On formule deux hypothèses contradictoires notées H_0 et H_1 dont on suppose que l'une et seulement l'une est vraie. Mathématiquement, formuler H_0 et H_1 revient à choisir deux sous-ensembles disjoints de Θ notés Θ_0 et de Θ_1 de sorte que l'hypothèse H_0 s'écrit alors $\{\theta_0 \in \Theta_0\}$ tandis que l'hypothèse H_1 s'écrit $\{\theta_1 \in \Theta_1\}$. Soit $(X_1, ..., X_n)$ un échantillon i.i.d. issu d'une variable parente X et soit $(x_1, ..., x_n) \in \mathcal{X}^n$ une réalisation de $(X_1, ..., X_n)$. Construire un test de H_0 contre H_1 revient à construire une région critique $\mathcal{R}$ de telle sorte que l'on rejette H_0 lorsque $(x_1, ..., x_n) \in \mathbb{R}$ et que l'on s'assure que le risque de 1ère espèce est inférieur ou égal à α.

Dans ce cadre, on parle de **fonction de risque de 1ère espèce** définie sur Θ_0 pour

$$\theta \to \alpha(\theta) = \mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R}) \,.$$

On appelle **taille du test** la probabilité maximale de rejeter H_0 alors que H_0 est vraie:

$$\sup_{\theta \in \Theta_0} \{\mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R})\} \,.$$

On dit qu'un test est de **niveau** α si sa taille est égale à α ie si

$$\sup_{\theta \in \Theta_0} \{\mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R})\} = \alpha \,.$$

On parle de **fonction de risque de 2nde espèce** définie sur Θ_1 pour

$$\theta \to \beta(\theta) = \mathbb{P}_\theta((X_1, ..., X_n) \notin \mathcal{R}) \,.$$

On parle de **fonction puissance** définie sur Θ_1 pour

$$\theta \to 1 - \beta(\theta) = \mathbb{P}_\theta((X_1, ..., X_n) \in \mathcal{R}) \,.$$

• **Introduction à la régression**:
Le but de tout modèle de régression (d'une manière générale) est de spécifier une relation (de nature stochastique) entre une variable Y appelée variable à expliquer (ou variable de réponse ou variable endogène ou variable dépendante) et un certain nombre de variables explicatives $X^{(1)}, ..., X^{(p)}$ (ou variables de contrôle ou variables exogènes ou régresseurs ou covariables). Ici, on supposera que les covariables sont de loi continue.
Notons $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ pour $i = 1, ..., n$ les observations recueillies sur n individus que l'on suppose distribuées comme $(Y, X^{(1)}, ..., X^{(p)})$. Le modèle de régression linéaire gaussien standard à n observations indépendantes s'écrit sous la forme générique suivante:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)} + \varepsilon_i, \qquad \text{où} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Le modèle de régression linéaire gaussien standard repose donc sur les hypothèses suivantes:

($\mathbf{M_1}$) normalité: conditionnellement à $X_i^{(1)}, ..., X_i^{(p)}$, la variable de réponse Y_i suit la loi normale.

($\mathbf{M_2}$) homoscédasticité: les lois conditionnelles de Y_i sachant $X_i^{(1)}, ..., X_i^{(p)}$ ont même variance donc pour $i = 1, ..., n$, on a $\sigma_i^2 = \sigma^2$.

($\mathbf{M_3}$) linéarité: le prédicteur linéaire $\eta_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)}$ est une combinaison affine des covariables, la linéarité s'apprécie relativement à la linéarité de η_i en les paramètres. Attention, la linéarité du modèle ne doit pas prêter à confusion: un modèle dit linéaire est linéaire en les paramètres β_0, β_1,...β_p. Ainsi, le modèle de régression linéaire simple

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

est, comme son nom l'indique, linéaire, tout comme l'est le modèle de régression linéaire multiple ($p > 1$):

$$Y = \beta_0 + \sum_{k=1}^{p} \beta_k X^{(k)} + \varepsilon\,.$$

Le modèle de régression polynomiale

$$Y = \beta_0 + \beta_1 X^{(1)} + \beta_2 (X^{(1)})^2 + \varepsilon$$

est un modèle linéaire bien que la relation entre Y et $X^{(1)}$ soit quadratique. En revanche, un modèle de la forme

$$Y = \beta_0 + \beta_1 \exp(\beta_2 X^{(1)} + \beta_3 X^{(2)}) + \varepsilon_i$$

ou bien de la forme

$$Y = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} + \varepsilon$$

n'est pas linéaire!

($\mathbf{M_4}$) centrage des erreurs: $\mathbb{E}[\varepsilon_i] = 0$ pour $i = 1, ..., n$.
Remarquons que cela équivaut à l'existence d'une relation "identité" entre le prédicteur linéaire η_i d'une part et l'espérance conditionnelle de la variable réponse d'autre part:

$$\mu_i := \mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \eta_i := \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)}.$$

($\mathbf{M_5}$) non-corrélation: les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont non-corrélés pour $i = 1, ..., n$. Si la normalité est effectivement respectée, cela revient alors à dire que les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont mutuellement indépendants.

($\mathbf{M_6}$) exogénéité = non-endogénéité: $(X_i^{(1)}, ..., X_i^{(p)})$ est indépendant de ε_i, pour $i = 1, ..., n$.

($\mathbf{M_7}$) non-colinéarité des covariables: les covariables sont non-corrélées entre elles, ce qui se traduit en pratique par le fait que les vecteurs $(X_1^{(k)}, ..., X_n^{(k)})$ sont non-colinéaires pour $k = 1, ...p$. On écartera également le cas trivial où l'un de ces vecteurs serait colinéaire au vecteur de longueur n dont toutes les composantes sont égales à 1.

• **Test de Student sur l'effet des covariables**:
Dans le cadre du modèle linéaire gaussien standard présenté ci-dessus, on se demande si les covariables introduites dans le modèle ont un effet statistiquement significatif ou non. Autrement dit, on souhaite tester au niveau de risque α l'hypothèse nulle H_0: $\beta_j = 0$ contre l'hypothèse alternative H_1: $\beta_j \neq 0$, ceci pour un $j \in \{1, \ldots, p\}$ fixé quelconque. Pour cela, on s'intéressera ici au test de Student qui est fondé sur l'estimateur des moindres carrés ordinaires (qui coïncide ici avec l'EMV) de β_j que l'on notera $\widehat{\beta}_j$. Notons $\widehat{\mathrm{Var}}\left(\widehat{\beta}_j\right)$ un etimateur de la variance de $\widehat{\beta}_j$. La statistique de test de Student pour le test de H_0: $\beta_j = 0$ s'écrit:

$$T_n^{(j)} = \frac{\widehat{\beta}_j}{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\beta}_j\right)}}.$$

Sous H_0, la loi de T_n est une loi de Student à $(n - (p + 1)$ degrés de liberté.
Le logiciel `R` permet d'effectuer simplement ces calculs. Pour ajuster un modèle linéaire sur un échantillon $(Y_i, X_i^{(1)}, ..., X_i^{(p)})_{i=1,...,n}$, on stockera les n valeurs des Y_i dans un vecteur `y`, puis, pour $k = 1, \ldots, p$ on stockera les n valeurs des $X_i^{(k)}$ dans un vecteur `xk`. Le modèle linéaire est alors ajusté au moyen de l'instruction suivante (où $p = 3$)

```
mylm<- lm(y~x1+x2+x3)
```

et le résultat est stocké dans l'objet `mylm`. Pour accéder à l'ensemble des quantités calculées, on peut exécuter l'instruction suivante:

```
summary(mylm)
```

Sont notamment calculées, les estimations des β_j pour $j = 0, \ldots, p$, les estimations de leurs écarts-types, la statistique de test de Student correspondante et la p-valeur du test associé. L'objet `summary(mylm)` est une liste dont on peut récupérer chacune des composantes qui nous intéresse. Pour cela, l'instruction `str(summary(myfit))` permet de voir le nom et la structure des différentes composantes de cette liste.

Exercice 1.

1. Illustrer de manière empirique à partir de données simulées le comportement de la taille du test.

2. Illustrer de manière empirique à partir de données simulées le comportement de la fonction puissance.

Vous veillerez à faire varier tour à tour:

- le risque de 1ère espèce α,
- la taille de l'échantillon n,
- le cas échéant l'ampleur de l'écart à H_0.

Que pensez-vous de la robustesse du test de Student à la bonne spécification du modèle (cf les hypothèses $(\mathbf{M_1})$-$(\mathbf{M_7})$)?
NB: on dit que le modèle est bien spécifié lorsqu'il correspond effectivement au mécanisme ayant servi à généré les données.