Apprentissage en grande dimension (Partie #2)

Thomas Verdebout

Université de Lille

Contenu du cours

1. Introduction.
2. Apprentissage supervisé: régression.
3. Apprentissage supervisé: classification.
4. Apprentissage non-supervisé

2: Apprentissage supervisé: régression: MCO.

Nous considérons dans cette partie des méthodes d'apprentissage supervisé paramétriques. Plus précisément, nous supposons observé $(\mathbf{X}_1, Y_1), \ldots (\mathbf{X}_n, Y_n)$ tels que

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip} + \epsilon_i = \boldsymbol{\beta}' \mathbf{X}_i + \epsilon_i,$$

où ϵ_i est un terme d'erreur aléatoire et où $\mathbf{X}_i := (1, X_{i1}, \ldots, X_{ip})'$ et $\boldsymbol{\beta} := (\beta_0, \beta_1, \ldots, \beta_p)'$. En posant $\mathbf{Y} := (Y_1, \ldots, Y_n)$, $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ et $\mathbf{X} := (\mathbf{X}_1, \ldots, \mathbf{X}_n)'$, nous pouvons réécrire le modèle de façon matricielle

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

avec $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ et (souvent) $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$.

Supposons que $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$. Alors la fonction de vraisemblance est donnée par

$$\begin{aligned}\ell(\boldsymbol{\mu}, \sigma | \mathbf{Y}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}(\mathbf{Y}-\boldsymbol{\mu})'(\mathbf{Y}-\boldsymbol{\mu})\right) \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Y}-\boldsymbol{\mu}\|^2\right).\end{aligned}$$

Si le modèle linéaire tient pour $Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$ alors l'estimateur maximum de vraisemblance gaussien pour (β, σ^2) équivaut à

$$(\hat{\beta}, \hat{\sigma}^2) = \operatorname{argmax}_{\beta, \sigma^2} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\beta\|^2\right).$$

Notons que

$$(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) = \text{argmax}_{\boldsymbol{\beta},\sigma^2} \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right)$$
$$= \text{argmin}_{\boldsymbol{\beta},\sigma^2} \left(\frac{n}{2} \log(\sigma^2) + \frac{1}{2\sigma^2} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \right).$$

Si nous fixons σ^2 pour le moment, alors nous devons trouver

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

En posant $\partial \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 / \partial \boldsymbol{\beta} = \mathbf{0}$, on obtient facilement (exercice) que

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Afin d'obtenir $\hat{\sigma}^2$ nous déterminons

$$\text{argmin}_{s>0} \left(\frac{n \log(s)}{2} + \frac{1}{2s} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 \right).$$

Des calculs directs (exercice) montrent que

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2.$$

Théorème

Supposons qu'on a un modèle linéaire de la forme $Y \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$. *Alors l'estimateur maximum de vraisemblance est donné par*

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \textit{et} \quad \hat{\sigma}^2 = \frac{1}{n} \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2.$$

Remarque. Les estimateurs

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad \text{et} \quad \hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2$$

ne sont pas sans intérêt quand $\mathbf{Y}$ n'est pas normalement distribué. L'estimateur $\hat{\beta}$ est aussi appelé *estimateur des moindres carrés.*

Le vecteur $\hat{\varepsilon} := \mathbf{Y} - \mathbf{X}\hat{\beta}$ est appelé *vecteur des résidus*. Nous notons

$$\begin{aligned} SCR &= \|\mathbf{Y} - \mathbf{X}\hat{\beta}\|^2 \\ &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \end{aligned}$$

la somme des carrés des résidus.

Une autre manière d'obtenir le MLE pour β consiste à écrire

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\mu} \in L} \|\mathbf{Y} - \boldsymbol{\mu}\|^2,$$

où L est l'espace linéaire de $\mathbb{R}^n$ engendré par les colonnes $\mathbf{X}$. Par définition $\hat{\boldsymbol{\mu}} = p_L(\mathbf{Y})$, la *projection* de $\mathbf{Y}$ sur L.

Notons que la projection $p_L(\mathbf{Y})$ est caractérisée comme l'unique élément dans L tel que

$$\langle \mathbf{Y} - p_L(\mathbf{Y}), \mathbf{x} \rangle = 0 \quad \text{pour tout} \quad \mathbf{x} \in L$$

$$\Longleftrightarrow$$

$$\mathbf{X}'(\mathbf{Y} - p_L(\mathbf{Y})) = \mathbf{0}.$$

Il est clair que $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \in L$ et on voit facilement que

$$\mathbf{X}'(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = \mathbf{0}.$$

D'où nous avons que

$$\hat{\boldsymbol{\mu}} = p_L(\mathbf{Y}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}}.$$

La matrice

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

est appelée *matrice de projection*. Elle est (exercice)

- symétrique: $\mathbf{H} = \mathbf{H}'$, et
- idempotente: $\mathbf{H}^2 = \mathbf{H}$.

La proposition suivante tient sans hypothèse de normalité.

Proposition

Supposons que $\mathbf{Y}$ *est un modèle linéaire de dimension p de la forme*

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim (\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Soient $\hat{\boldsymbol{\mu}}$, $\hat{\beta}$ *et* $\hat{\sigma}^2$ *définis comme avant. Alors*

$$E\hat{\beta} = \beta \quad \textit{et} \quad E\hat{\boldsymbol{\mu}} = \boldsymbol{\mu};$$
$$\mathrm{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad \textit{et} \quad \mathrm{Var}(\hat{\boldsymbol{\mu}}) = \sigma^2\mathbf{H};$$
$$E\hat{\sigma}^2 = \frac{n-p-1}{n}\sigma^2.$$

Preuve. La moyenne et la variance de $\hat{\beta}$ et $\hat{\mu}$ sont simples (tableau). Nous allons calculer l'espérance de $\hat{\sigma}^2$. En ayant recours à $\mathbf{H}' = \mathbf{H}$ et $\mathbf{H}^2 = \mathbf{H}$, nous obtenons que la même chose est vraie pour $\mathbf{I}_n - \mathbf{H}$. Notez aussi que $\mathbf{H}\boldsymbol{\mu} = \boldsymbol{\mu}$ et donc $(\mathbf{I}_n - \mathbf{H})\boldsymbol{\mu} = \mathbf{0}$. Par conséquent

$$\begin{aligned} E\hat{\sigma}^2 &= \frac{1}{n} E\|(\mathbf{I}_n - \mathbf{H})\mathbf{Y}\|^2 = \frac{1}{n} E\|(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}\|^2 \\ &= \frac{1}{n} E\mathrm{tr}(\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})'(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}) \\ &= \frac{1}{n}\mathrm{tr}(E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{I}_n - \mathbf{H})) = \frac{\sigma^2}{n}\mathrm{tr}(\mathbf{I}_n - \mathbf{H}). \end{aligned}$$

La preuve suit alors du fait que
$\mathrm{tr}(\mathbf{H}) = \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathrm{tr}(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) = p + 1.$ □

Proposition

Soit $\mathbf{Y}$ *tel que*

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n).$$

Soient $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\beta}}$ *et* $\hat{\sigma}^2$ *définis comme avant. Alors*

(i) $\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$;

(ii) $\hat{\mu} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \mathbf{H})$;

(iii) $\hat{\sigma}^2 \sim \dfrac{\sigma^2}{n} \chi^2_{n-p-1}$;

(iv) $\hat{\sigma}^2$ *et* $\hat{\boldsymbol{\beta}}$ *(ou* $\hat{\boldsymbol{\mu}}$*, respectivement) sont indépendants.*

Preuve de (i) et (ii). Les vecteurs $\hat{\beta}$ et $\hat{\boldsymbol{\mu}}$ sont des transformations linéaires du vecteur normal $\mathbf{Y}$. Ils sont donc gaussiens. Par la proposition précédente, nous connaissons leur moyenne et variance. Nous obtenons donc ainsi (i) et (ii). □

Pour prouver (iii) et (iv), nous allons utiliser quelques résultats d'algèbre linéaire.

Sous-espaces orthogonaux. Deux sous-espaces linéaires L_1 et L_2 de $\mathbb{R}^n$ sont appelés orthogonaux si $\mathbf{x} \perp \mathbf{y}$ pour tout $\mathbf{x} \in L_1$ et $\mathbf{y} \in L_2$. Alors nous définissons

$$L_1 \oplus L_2 = \{\mathbf{x} + \mathbf{y} | \mathbf{x} \in L_1 \text{ et } \mathbf{y} \in L_2\}.$$

$$L_1 \ominus L_2 = \{\mathbf{x} \in L_1 | \mathbf{x} \perp L_2\} \quad \text{(en supposant que } L_2 \subset L_1\text{)}.$$

Lemma

Supposons que $L_2 \subset L_1$ sont des sous-espaces linéaires de $\mathbb{R}^n$. Alors

(i) $L_1 \ominus L_2$ *est une sous-espace linéaire*;

(ii) $L_1 \ominus L_2 \perp L_2$;

(iii) $(L_1 \ominus L_2) \oplus L_2 = L_1 \quad (\Longrightarrow L^\perp \oplus L = \mathbb{R}^n)$;

(iv) $p_{L^\perp}(\mathbf{v}) = \mathbf{v} - p_L(\mathbf{v})$.

Proposition

Soient $L_1, L_2, \ldots, L_r$ *des sous-espaces orthogonaux de* $\mathbb{R}^n$ *avec* $\dim(L_i) = p_i$ *et* $L_1 \oplus L_2 \oplus \cdots \oplus L_r = \mathbb{R}^n$ *et soit* $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I}_n)$. *Alors*

$$(a) \quad p_{L_1}(\varepsilon), \ldots, p_{L_r}(\varepsilon) \quad \textit{sont indépendants.}$$

$$(b) \quad \|p_{L_i}(\varepsilon)\|^2 \sim \sigma^2 \chi^2_{p_i} \quad 1 \leq i \leq r.$$

Preuve de (a). Comme $L_1, L_2, \ldots, L_r$ sont des sous-espaces orthogonaux de $\mathbb{R}^n$ avec $\dim(L_i) = p_i$ et $L_1 \oplus L_2 \oplus \cdots \oplus L_r = \mathbb{R}^n$, nous avons une base orthonormale (BON)

$$\{\mathbf{e}_{ij},\ 1 \leq j \leq p_i,\ 1 \leq i \leq r\}$$

de $\mathbb{R}^n$ tel que $\mathbf{E}_i = \{\mathbf{e}_{ij},\ 1 \leq j \leq p_i\}$ est une BON de L_i. Puisque $\mathbf{E}_i'\mathbf{E}_i = \mathbf{I}_{p_i}$, nous avons que

$$p_{L_i}(\varepsilon) = \mathbf{E}_i(\mathbf{E}_i'\mathbf{E}_i)^{-1}\mathbf{E}_i'\varepsilon = \mathbf{E}_i\mathbf{E}_i'\varepsilon.$$

Donc, comme $\mathbf{E}_i'\mathbf{E}_j = 0$ pour $i \neq j$, nous obtenons

$$\mathrm{Cov}(p_{L_i}(\varepsilon), p_{L_j}(\varepsilon)) = \mathbf{E}_i\mathbf{E}_i'\mathrm{Cov}(\varepsilon, \varepsilon)\mathbf{E}_j\mathbf{E}_j' = 0.$$

Preuve de (b). Puisque $\{\mathbf{e}_{ij},\ 1 \leq j \leq p_i,\ 1 \leq i \leq r\}$ est une BON de $\mathbb{R}^n$, on a que

$$\begin{aligned} \|p_{L_i}(\varepsilon)\|^2 &= \|\sum_{j=1}^{p_i} \langle \mathbf{e}_{ij}, \varepsilon \rangle \mathbf{e}_{ij}\|^2 \\ &= \sum_{j=1}^{p_i} \langle \mathbf{e}_{ij}, \varepsilon \rangle^2. \end{aligned}$$

Puisque les e_{ij} sont orthonormaux, il suit que les variables $\langle \mathbf{e}_{ij}, \varepsilon \rangle$ sont des variables aléatoires indépendantes de loi normale. Leur moyenne est 0 et la variance σ^2. □

Preuve de (iii) et (iv) de la proposition précédente. Rappelons que

$$\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2.$$

Nous rappelons que

$$\|\mathbf{Y} - \hat{\boldsymbol{\mu}}\|^2 = \|(\mathbf{I}_n - \mathbf{H})\boldsymbol{\varepsilon}\|^2.$$

Par le point (iv) du lemme précédent il découle que $(\mathbf{I}_n - \mathbf{H})\mathbf{v} = p_{L^\perp}(\mathbf{v})$, la projection de $\mathbf{v}$ sur l'espace

$$L^\perp = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x} \perp L = \mathrm{span}(\mathbf{X})\}.$$

Comme $\dim(L^\perp) = n - p - 1$, la preuve de (iii) suit du point (b) de la proposition précédente.

Comme

$$\hat{\boldsymbol{\mu}} = \mathbf{HY} = \mathbf{X}\beta + \mathbf{H}\varepsilon = \mathbf{X}\beta + p_L(\varepsilon)$$

et

$$\hat{\sigma}^2 = \frac{1}{n}\|(\mathbf{I}_n - \mathbf{H})\varepsilon\|^2 = \frac{1}{n}\|p_{L^\perp}(\varepsilon)\|^2,$$

nous concluons du point (a) de la proposition précédente que $\hat{\boldsymbol{\mu}}$ et $\hat{\sigma}^2$ sont indépendants.

L'indpendance de $\hat{\beta}$ et $\hat{\sigma}^2$ decoule de

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\boldsymbol{\mu}}.$$

□

Pour un modèle linéaire $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$, nous allons tester des hypothèses de la forme suivante:

$$H_0 : \boldsymbol{\mu} \in L_2 \quad \text{contre} \quad H_1 : \boldsymbol{\mu} \in L_1 \backslash L_2 \quad (L_2 \subset L_1).$$

Nous supposons dans la suite que $\dim(L_1) = p_1$ et $\dim(L_2) = p_2 \leq p_1$. La statistique du test de rapport de vraisemblance est donnée par

$$\Lambda(\mathbf{Y}) = \left(\frac{\|\mathbf{Y} - p_{L_1}(\mathbf{Y})\|^2}{\|\mathbf{Y} - p_{L_2}(\mathbf{Y})\|^2} \right)^{n/2},$$

et rejette l'hypothèse nulle si $\Lambda(\mathbf{Y})$ est trop petit.

Maintenant comme

$$\|\mathbf{Y} - p_{L_2}(\mathbf{Y})\|^2 = \|\mathbf{Y} - p_{L_1}(\mathbf{Y}) + \underbrace{p_{L_1}(\mathbf{Y}) - p_{L_2}(\mathbf{Y})}_{\in L_1}\|^2$$

nous obtenons par le théorème de projection et le théorème de Pythagore que

$$\|\mathbf{Y} - p_{L_2}(\mathbf{Y})\|^2 = \|\mathbf{Y} - p_{L_1}(\mathbf{Y})\|^2 + \|p_{L_1}(\mathbf{Y}) - p_{L_2}(\mathbf{Y})\|^2.$$

Donc le test de rapport de vraisemblance est équivalent au test qui rejette l'hypothèse nulle si

$$F(\mathbf{Y}) = \frac{\|p_{L_1}(\mathbf{Y}) - p_{L_2}(\mathbf{Y})\|^2/(p_1 - p_2)}{\|\mathbf{Y} - p_{L_1}(\mathbf{Y})\|^2/(n - p_1)}$$

est trop grand.

Maintenant sous H_0 : $\boldsymbol{\mu} \in L_2$ nous avons

$$
\begin{aligned}
F(Y) &= \frac{\|p_{L_1}(\mathbf{Y}) - p_{L_2}(\mathbf{Y})\|^2/(p_1 - p_2)}{\|\mathbf{Y} - p_{L_1}(\mathbf{Y})\|^2/(n - p_1)} \\
&= \frac{\|p_{L_1}(\varepsilon) - p_{L_2}(\varepsilon)\|^2/(p_1 - p_2)}{\|\varepsilon - p_{L_1}(\varepsilon)\|^2/(n - p_1)} \\
&= \frac{\|p_{L_1 \ominus L_2}(\varepsilon)\|^2/(p_1 - p_2)}{\|p_{L_1^\perp}(\varepsilon)\|^2/(n - p_1)}.
\end{aligned}
$$

Comme $\mathbb{R}^n = L_1^\perp \oplus (L_1 \ominus L_2) \oplus L_2$ nous concluons que

$$
\|p_{L_1 \ominus L_2}(\varepsilon)\|^2 \sim \chi^2_{p_1 - p_2} \quad \perp\!\!\!\perp \quad \|p_{L_1^\perp}(\varepsilon)\|^2 \sim \chi^2_{n - p_1}.
$$

Théorème

Supposons qu'on a un modèle linéaire de la forme $\mathbf{Y} \sim \mathcal{N}_n(\boldsymbol{\mu}, \sigma^2 I_n)$. *Le test de rapport de vraisemblance pour* $H_0 : \mu \in L_2$ *contre* $H_1 : \mu \in L_1$ *est équivalent à un test F. La statistique de test*

$$F(\mathbf{Y}) = \frac{\|p_{L_1}(\mathbf{Y}) - p_{L_2}(\mathbf{Y})\|^2/(p_1 - p_2)}{\|\mathbf{Y} - p_{L_1}(\mathbf{Y})\|^2/(n - p_1)}$$

suit une distribution de Fisher avec $p_1 - p_2$ *et* $n - p_1$ *degrés de liberté. Pour un test de niveau* α, *nous rejetons* H_0 *si*

$$F(\mathbf{Y}) \geq F_{p_1-p_2,n-p_1;1-\alpha}.$$

Exemple: test de signification globale du modèle. Les hypothèses du test:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ \\ H_1 : \exists j \in \{1, \ldots, p\} \text{ tq } \beta_j \neq 0. \end{cases}$$

Clairement pour ce problème, $\dim(L_2) = 1$ et $\dim(L_1) = p + 1$. On obtient ici

$$F(\mathbf{Y}) = \frac{SCE/p}{SCR/n - p - 1},$$

où $SCE = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ et pour rappel $SCR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$. Pour rappel,

$$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = SCE + SCR.$$

On pourrait aussi vouloir tester l'hypothèse nulle $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$. Ceci peut se faire avec un test F comme décrit précédemment.

On peut tester plusieurs hypothèses de ce type en faisant varier q, etc et notamment obtenir des t-tests pour la significativité des variables individuelles.

Étant donné les p-valeurs associés aux t-tests pour chaque variable, pourquoi avons-nous besoin d'examiner la statistique F globale? Après tout, il semble probable que si l'une des p-valeurs pour les variables individuelles est très faible, alors au moins un des prédicteurs est lié à la réponse. Toutefois, cette logique est erronée, surtout lorsque le nombre de covariables p est élevé.

Par exemple, supposons que $p = 200$ et que l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$ est vraie. Dans cette situation, il est probable qu'au moins une *p*-valeur soit inférieure à .05. Par conséquent, si nous utilisons les statistiques *t* individuelles et les *p*-valeurs associées pour décider s'il existe ou non une association entre les variables et la réponse, il y a de fortes chances que nous concluions à tort qu'il existe une relation. Cependant, la statistique *F* globale ne souffre pas de ce problème car elle s'ajuste au nombre de prédicteurs. Par conséquent, si l'hypothèse nulle est vraie, il n'y a que 5 pourcent de chances que la statistique F donne une *p*-valeur inférieure à .05, quel que soit le nombre de prédicteurs ou le nombre d'observations.

- Généralement a première étape d'une analyse de régression multiple consiste à calculer la statistique F (pour tout le modèle) et à examiner la p-valeur associée. Si nous concluons, sur la base de cette p-valeur, qu'au moins un des prédicteurs est lié à la réponse, il est naturel de se demander quels sont les coupables !
- La tâche consistant à déterminer quels prédicteurs sont associés à la réponse, afin d'ajuster un modèle unique impliquant uniquement ces prédicteurs, est appelée sélection de variables.

Idéalement, nous aimerions effectuer une sélection de variables en essayant un grand nombre de modèles différents, chacun contenant un sous-ensemble différent de prédicteurs. différents modèles, chacun contenant un sous-ensemble différent de prédicteurs. Par exemple, si $p = 2$, nous pouvons envisager quatre modèles : (1) un modèle ne contenant aucune variable, (2) un modèle ne contenant que X_1, (3) un modèle ne contenant que X_2 et (4) un modèle contenant à la fois X_1 et X_2. Nous pouvons alors sélectionner le meilleur modèle parmi tous les modèles que nous avons considérés. Comment déterminons-nous le meilleur modèle ? Diverses statistiques peuvent être utilisées pour pour juger de la qualité d'un modèle (c.f. TD)

Si $p = 30$...Les méthodes classiques sont:

- Forward selection: Nous commençons par un modèle qui contient une ordonnée à l'origine mais pas de prédicteurs. On réalise ensuite p régressions linéaires simples et ajoutons au modèle nul la variable qui donne la SCR la plus basse. Nous ajoutons ensuite à ce modèle la variable qui entraîne la SCR la plus faible avec deux variables (dont la première sélectionnée), etc. On utilise une "stopping rule"
- Backward selection: Nous commençons par toutes les variables du modèle et supprimons la variable ayant la p-valeur la plus élevée, c'est-à-dire la variable potentiellement la moins significative. Le nouveau modèle à $(p - 1)$ variables est ajusté, et la variable ayant la plus grande p-valeur est supprimée. Cette procédure se poursuit jusqu'à ce qu'une règle d'arrêt soit atteinte.

- le modèle contenant tous les prédicteurs aura toujours la plus petite SCR et le plus grand $R^2 := 1 - \frac{SCR}{SCT}$
- On a donc le R^2 ajusté $R^2_{aju} := 1 - \frac{SCR/(n-p-1)}{SCT(n-1)}$
- Mais aussi...Mallow's

$$C_p := \frac{1}{n}(SCR + 2(p+1)\hat{\sigma}^2)$$

, où $\hat{\sigma}^2$ est un estimateur de la variance du modèle. On sélectionne le modèle avec le C_p le plus petit

- Mais aussi AIC, BIC, etc
- C_p, AIC et BIC ont tous des justifications théoriques rigoureuses qui dépassent le cadre de ce cours. Toutes ces mesures sont simples à utiliser et à calculer.

Exemples de librairies pour la régression linéaire:

```
library(MASS)
library(ISLR2)
lm.fit <- lm(medv ~ lstat, data = Boston)
summary(lm.fit)
confint(lm.fit)
###
predict(lm.fit, data.frame(lstat = (c(5, 10, 15))),
    interval = "confidence")
predict(lm.fit, data.frame(lstat = (c(5, 10, 15))),
    interval = "prediction")
```

```
plot(lstat, medv)
abline(lm.fit)
###
abline(lm.fit, lwd = 3)
abline(lm.fit, lwd = 3, col = "red")
plot(lstat, medv, col = "red")
plot(lstat, medv, pch = 20)
plot(lstat, medv, pch = "+")
plot(1:20, 1:20, pch = 1:20)
###
par(mfrow = c(2, 2))
plot(lm.fit)
###
plot(predict(lm.fit), residuals(lm.fit))
plot(predict(lm.fit), rstudent(lm.fit))
###
plot(hatvalues(lm.fit))
which.max(hatvalues(lm.fit))
```

Régression multiple:

```
lm.fit <- lm(medv ~ lstat + age, data = Boston)
summary(lm.fit)
lm.fit <- lm(medv ~ ., data = Boston)
summary(lm.fit)
```