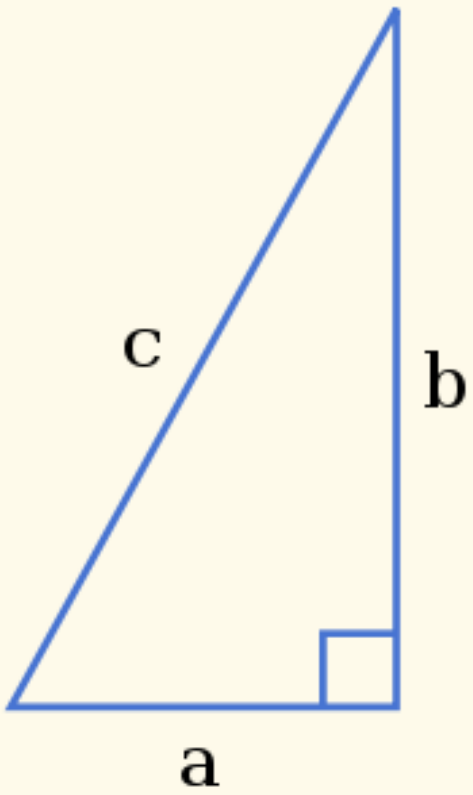
GC – Probabilités et Statistique

http://moodle.epfl.ch/course/view.php?id=14271

Cours 10b

- Géometrie de régression
- Logiciel R / interprétion des sorties R

Théorème de Pythagore



$$a^2 + b^2 = c^2$$

Géometrie de moindres carrés

- On considère $\boldsymbol{y}$ comme un vecteur dans l'espace n-dim
- Les vecteurs des colonnes de $\boldsymbol{X}$ forment un sous-espace (de l'estimation ou du modèle) p-dim
  - Variation des valeurs estimées des coefficients de régression localise des points différents du sous-espace
- Les valeurs prédites $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$ représentent le point du sous-espace le plus proche des observations : MCO est la *projection orthogonale* de $\boldsymbol{y}$ sur le sous-espace de $\boldsymbol{X}$
- Le résidu $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ est *orthogonal* aux vecteurs du sous-espace
- $SCE = \sum e_i^2 = \boldsymbol{e}'\boldsymbol{e}$ est le carré de la distance du vecteur des obs. au point le plus proche dans le sous-espace
- Partition de $\boldsymbol{y}$ en *deux composantes orthogonales* :
  - $\hat{\boldsymbol{y}}$ (sous-space du modèle, p dims)
  - $\hat{\boldsymbol{y}} - \boldsymbol{y}$ (sous-espace de l'erreur, $n - p$ dims)
- (degrés de liberté correspondent aux dims des sous-espaces)

Géométrie de MC

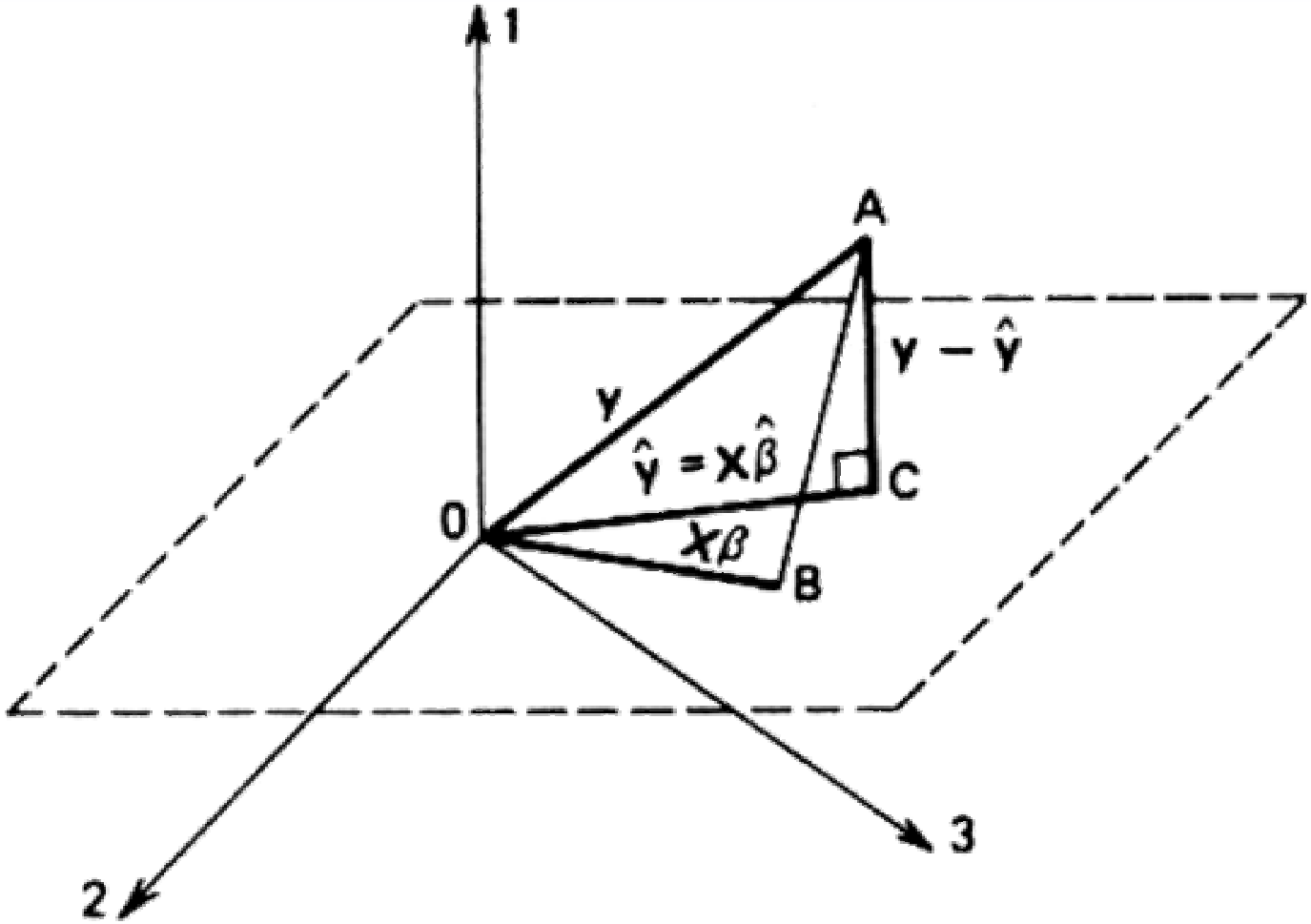


Figure 4.2 A geometrical interpretation of least squares.

Tableau de l'analyse de variance (ANOVA)

- Il s'agit d'une *partition de la somme des carrés totaux (SCT)*
- Théorème de Pythagore :

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- également :

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Cette égalité présentée dans un tableau :

Tableau d'ANOVA

source	df	SC (SS)	CM (MS) ($=SC/\text{df}$)	F	p-valeur
régression	p	$SCM = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	SCM/p	CMM/CME	$P(F_{obs} > F_{p,n-p-1})$
erreur	$n-p-1$	$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$SCE/(n-p-1) (= \hat{\sigma}^2)$		
total (corr.)	$n-1$	$SCT = \sum_{i=1}^{n} (y_i - \overline{y})^2$			

F-test - régression

- La statistique $F_{obs} = CM(\text{source})/CME$ teste l'hypothèse $H_0 : \beta_1 = \ldots = \beta_p = 0$ vs. A : au moins 1 $\beta_i \neq 0$
- La distribution de F_{obs} si H est vraie est *la distribution* $F_{p,n-p-1}$ de Fisher
- Au numérateur de la statistique F_{obs} se trouve *la variance expliquée par le modèle de régression*
- Au dénominateur se trouve *la variance résiduelle*
- On REJETTE l'hypothèse nulle H pour *grandes valeurs de F*
- Lorsqu'on teste une seule pente ($H : \beta_i = 0$), $F_{1,n} = t_n^2$

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

$F_{p,n-p-1}$

p-valeur

Coefficient de détermination

- La valeur y_i d'une observation peut être décomposée en deux parties : une partie *expliquée par le modèle* et une partie *résiduelle*
- La dispersion de l'ensemble des observations se décompose donc en :
  1. variance expliquée par la régression, et
  2. variance résiduelle, inexpliquée
- Le *coefficient de détermination* (ou *corrélation multiple*) R^2 se définit alors comme la part de variance expliquée par rapport à la variance totale
- Également, $R^2 = 1 - SCE/SCT$
- Dans le cadre d'une régression linéaire *simple*, c'est *le carré du coefficient de corrélation*

Coefficient de détermination ajusté

- Le *coefficient de détermination ajusté* R^2_{aj} tient compte du *nombre de variables*
- En effet, le défaut principal du R^2 est de *croître avec le nombre de variables explicatives*
- Un excès de variables produit des modèles *peu robustes*
- Donc on s'intéresse davantage à cet indicateur (R^2_{aj}) qu'au R^2
- Ce n'est pas vraiment un 'carré' – il peut même être négatif

$$R^2_{aj} = 1 - \frac{SCE/(n-p-1)}{SCT/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1}$$

L'estimation de régression

```
> trees.fit <- lm(Volume ~ Diameter + Height, trees.dat)
> summary(trees.fit)

Call:
lm(formula = Volume ~ Diameter + Height, data = trees.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

R^2

R^2-ajusté

R^2 ou R^2-ajusté ?

UTILISEZ LE R^2 AJUSTÉ !

MARRE DU R^2 ? Comme monsieur Statos, optez pour une qualité de régression plus sûre !!!

« Avant, j'utilisais un R^2 normal, j'étais fatigué et ça se voyait sur mon visage ; depuis que j'ai découvert le R^2 ajusté, ma vie a complètement changé ! »

Dépêchez-vous !

SATISFAIT ou REMBOURSÉ (*)

VU SUR INTERNET !!!

(*) voir conditions au verso

Dernière minute :

Pour vous souhaiter la bienvenue, la somme des carrés des résidus vous est offerte !

Tester un sous-modèle

- Modèle complet (Ω) : $y = \beta_0 + \beta_1 + \ldots + \beta_p$
- Sous-modèle (ω) : $y = \beta_0 + \beta_1 + \ldots + \beta_q$, $q < p$
- $H : \beta_{q+1} = \cdots = \beta_p = 0$ vs. A : au moins 1 $\beta_i \neq 0$, $q+1 \leq i \leq p$

Tableau d'ANOVA

source	df	SC (SS)	CM (MS) ($=SC/\text{df}$)
ω	q	$SCM(\omega)$	SCM/q
termes suppl.	$p-q$	$SCE(\omega) - SCE(\Omega)$	$(SCE(\omega) - SCE(\Omega))/(p-q)$
erreur	$n-p-1$	$SCE(\Omega)$	$SCE(\Omega)/(n-p-1)$
total (corr.)	$n-1$	SCT	

- La statistique F pour tester la signification des termes supplémentaires dans Ω est :

$$F_{obs} = \frac{(SCE(\omega) - SCE(\Omega))/(p-q)}{SCE(\Omega)/(n-p-1)} \sim F_{p-q,n-p-1} \text{ sous } H$$

- Donc on REJETTE H lorsque $F_{obs} > F_{p-q,n-p-1}(1-\alpha)$

Exemple 10b.1

Pour un échantillon aléatoire de communes, on a les données suivants :

- Y = pourcentage des adultes qui votent
- X_1 = pourcentage des adultes proprietaires
- X_2 = pourcentage des adultes personnes de couleur
- X_3 = revenu médiane de la famille (milliers CHF)
- X_4 = âge médiane
- X_5 = pourcentage des adultes résident au moins 10 années

Exemple 10b.1, cont.

	Sum of Squares	DF	Mean Square	F	Sig	R-Square
Regression	----	---	----	----	----	----
Residual	2940.0	---	----			Root MSE
Total	3753.3	---				----

Variable	Parameter Estimate	Standard Error	t	Sig
Intercept	70.0000			
x1	0.1000	0.0450	----	----
x2	-0.1500	0.0750	----	----
x3	0.1000	0.2000	----	----
x4	-0.0400	0.0500	----	----
x5	0.1200	0.0500	----	----

Exemple 10b.1, cont.– Activité

- Semble-t-il nécessaire d'inclure toute ces variables explicatrices dans le modèle ? Expliquer.

- La valeur F est utilisée pour quel test ? Interpreter le résultat de ce test.

- La valeur t de la variable X_1 est utilisée pour quel test ? Interpreter le résultat de ce test.

Exemple 10b.1, cont. – Activité

- Donner un IC à 95% pour le changement de la moyenne d'Y quand le pourcentage de proprietaires augmente par 1, en contrôlant pour les effets des autres variables ; l'interpreter.

- Donner un IC à 95% pour le changement de la moyenne d'Y quand le pourcentage de proprietaires augmente par 50, en contrôlant pour les effets des autres variables ; l'interpreter.