ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

MAP 433 : Statistique
Régression Linéaire

PC8

Quels métiers en Mathématiques Appliquées?

The worst and best jobs in 2014 (from CareerCast, 2014)

The Best	The Worst
1. Mathematician	200. Lumberjack
2. University Professor	199. Newspaper Reporter
3. Statistician	198. Enlisted Military Personnel
4. Actuary	197. Taxi Driver
5. Audiologist	196. Broadcaster

Un panorama partiel des metiers en Mathématiques Appliquées

Secteur	Exemples d'employeurs	Métiers
Business Analytics et optimisation de la production	Cabinets de consultants (Capgemini, Accenture, etc), PME / start-up (fifty-five, vertica, Mu Sigma, Eurodecision, MFG labs, Palantir, spatialytics, etc) la plupart des grands groupes (en interne)	· Marketing · Gestion des ressources (approvisionnement, ressources humaines) · Gestion des tarifications · Optimisation de la conception et des procédés · Recherche opérationnelle
Services web et logiciels	Services web (Google, Yahoo, etc), Logiciels (Microsoft, IBM, Dassault-system, Xerox, etc) SSII et start-up (IBM, Logica, CSC, GFI informatique, etc) Paiement électronique (Visa, E-commerce, etc)	· Moteurs de recherche, fonctionnalités web, etc · Logiciels génériques ou solutions spécialisées · Cryptographie (services sécurisés)
R&D réseaux et communication	Opérateurs mobile (Orange, Bouygues, Free, SFR, etc), Constructeurs (Alcatel-Lucent, Huawei, Ericsson, Sagem, etc),	· Planification réseaux · Prospective technologie et équipements · Qualité de service
Analyste statisticien	Industrie pharmaceutique (Sanofi, Servier, etc), Biointelligence, Toutes les branches de l'industrie / agroalimentaire, Organismes parapubliques (sécurité sanitaire, surveillance d'épidémie / pollution, services sociaux et de santé, etc)	· Biostatistiques · Production d'indices et prévision (trafic, consommation, ozone, coûts, marché, etc)
R&D signal & images	Thales, Safran, Dassault-system, Matra, General Electric, etc	· Traitement du signal et images · Guidage et contrôle · Imagerie médicale
R&D énergie, transport et	RTE, EDF, Areva, Veolia, SNCF, Schlumberger, Industrie pétrolière (Total, etc), Michelin, Renault, PSA, EADS, Dassault-	· Analyse, prévision, prospective · Gestion des risques · Modélisation, dimensionnement,

http://www.cmap.polytechnique.fr/~giraud/MetiersMaths.html

La régression linéaire: Théorie

Le modèle de régression

Observations

- **Réponse:** $y_1, \ldots, y_n \in \mathbb{R}$,
- **Covariables:** $x_1, \ldots, x_n \in \mathbb{R}^k$.

Modèle de régression

$$y_i = r(x_i) + \xi_i \quad \text{où } \mathbb{E}(\xi_i) = 0.$$

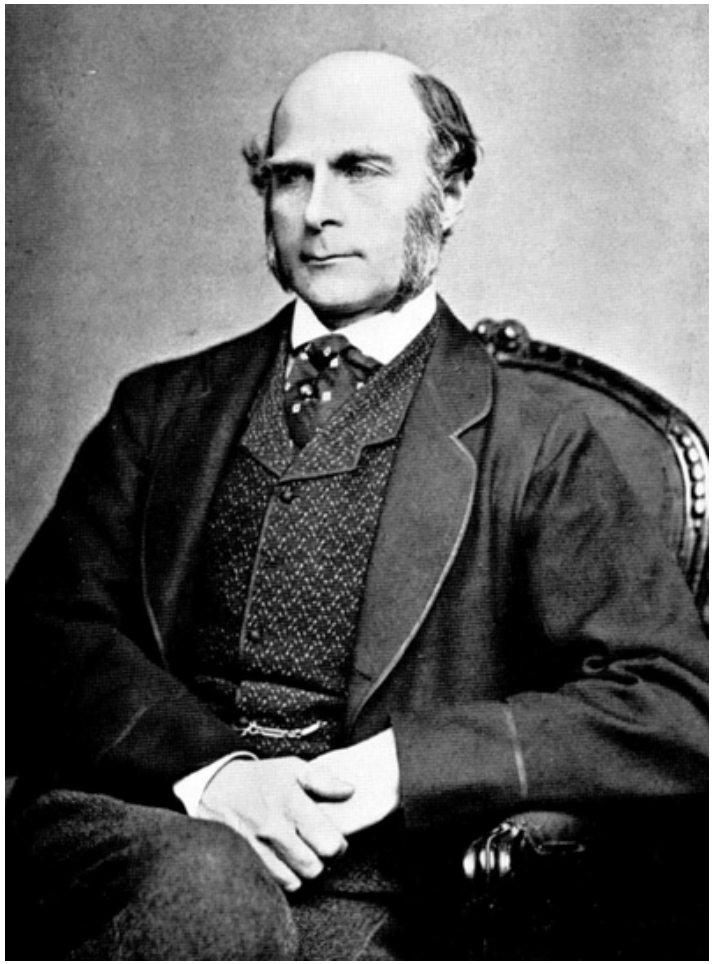
Régression Linéaire

$$y_i = x_i^T \theta + \xi_i \quad \text{où } \theta \in \mathbb{R}^k.$$

Francis Galton (1822–1911)

- Cousin de Darwin
- Mathématise la théorie de l'Evolution
- Précurseur en statistiques

mais.... fondateur du courant eugéniste ☹



Estimateur des moindres carrés

Modèle linéaire

Modèle: $r(x) = x^T\theta$ donc

$$Y = X\theta + \xi \quad \text{où } Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} \quad \text{et } \xi = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_n \end{bmatrix}.$$

Estimateur des moindres carrés

$$\hat{\theta}_{\mathrm{MC}} = \operatorname*{argmin}_{\theta \in \mathbb{R}^k} \|Y - X\theta\|^2$$

Formules:

$$X\hat{\theta}_{\mathrm{MC}} = \mathrm{Proj}_{\mathrm{Im}(X)}(Y) \quad \text{et} \quad \hat{\theta}_{\mathrm{MC}} = (X^TX)^{-1}X^TY.$$

Cas gaussien et design fixe

Si $\xi_i \sim \mathcal{N}(0, \sigma^2)$, x_i déterministes et $\hat{\theta}$ estimateur des MC on a:

$$\hat{\sigma}^2 = \frac{1}{n-k}\|Y - X\hat{\theta}\|^2 \quad \text{vérifie} \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

et

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}} \sim \text{Student}(n-k) := \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-k)}{n-k}}}.$$

Cas non gaussien: approximativement vrai pour n grand.

t-value: $\hat{T}_j = \hat{\theta}_j / \sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}$

Student?

Cas Gaussien

Cas gaussien et design fixe

Si $\xi_i \sim \mathcal{N}(0, \sigma^2)$, x_i déterministes et $\hat{\theta}$ estimateur des MC on a:

$$\hat{\sigma}^2 = \frac{1}{n-k}\|Y - X\hat{\theta}\|^2 \quad \text{vérifie} \quad \mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

et

$$\frac{\hat{\theta}_j - \theta_j}{\sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}} \sim \text{Student}(n-k) := \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-k)}{n-k}}}.$$

Cas non gaussien: approximativement vrai pour n grand.

t-value: $\hat{T}_j = \hat{\theta}_j / \sqrt{\hat{\sigma}^2[(X^TX)^{-1}]_{jj}}$

Student?

William Gosset (1876–1937)

- ingénieur agronome chez Guinness
- invente les $t - values$
- signe sous le nom de “Student”

Ronald Fisher (1890–1962)

- fondateur des statistiques classiques
- contributions majeurs en génétique des populations
- Livre majeur (1925) : *Statistical Methods for Research Workers*

La régression linéaire: un exemple

Logiciel d'analyse

Analyses réalisées avec R : c'est le logiciel standard en statistiques:

- gratuit! `http://cran.r-project.org/`
- partage de packages par toute la communauté stat :
  - principales analyses prêtes à l'emploi
  - accès aux procédures "état de l'art"
- défaut principal : aide en ligne peu performante. Pour apprendre : `http://cran.r-project.org/doc/manuals/R-intro.pdf`

Références

- Code R des analyses téléchargeable sur ma page web
- Livre: *Régression avec R* de Cornillon & Matzner-Lober

Données "ozone"

Relevés d'ozone (O3) et de covariables (Temperature, Nébulosité, Vent, etc) par l'association "Air Breizh".

```
> load("ozone.Rdata")
> dim(ozone)
50 10
> head(ozone)
```

O3	T12	T15	Ne12	N12	S12	E12	W12	Vx	O3v
63.60	13.40	15.00	7	0	0	3	0	9.35	95.60
89.60	15.00	15.70	4	3	0	0	0	5.40	100.20
79.00	7.90	10.10	8	0	0	7	0	19.30	105.60
81.20	13.10	11.70	7	7	0	0	0	12.60	95.20
88.00	14.10	16.00	6	0	0	0	6	-20.30	82.80
68.40	16.70	18.10	7	0	3	0	0	-3.69	71.40

Régression linéaire

```
> reg1 <- lm(O3~.,data=ozone)
> names(reg1)
"coefficients" "residuals" "effects" "rank" "fitted.values"
"assign" "qr" "df.residual" "xlevels" "call" "terms" "model"
> reg1$coefficients
```

variable	Estimator
(Intercept)	54.7278
T12	-0.3518
T15	1.4972
Ne12	-4.1922
N12	1.2755
S12	3.1711
E12	0.5277
W12	2.4749
Vx	0.6077
O3v	0.2454

Inspection des résidus (1/4)

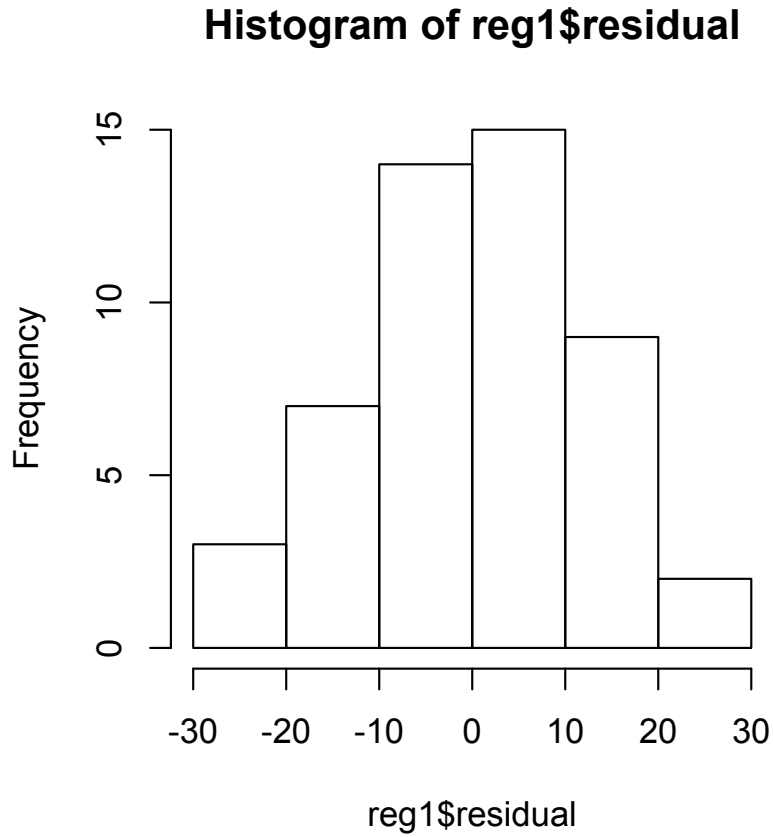
Distribution des résidus?

Les résidus $\hat{\xi} = Y - X\hat{\theta}$ sont donnés par

```
> reg1$residuals
```

On peut tracer un histogramme des résidus

```
> hist(reg1$residuals)
```



Histogramme des résidus

Inspection des résidus (2/4)

Quantiles

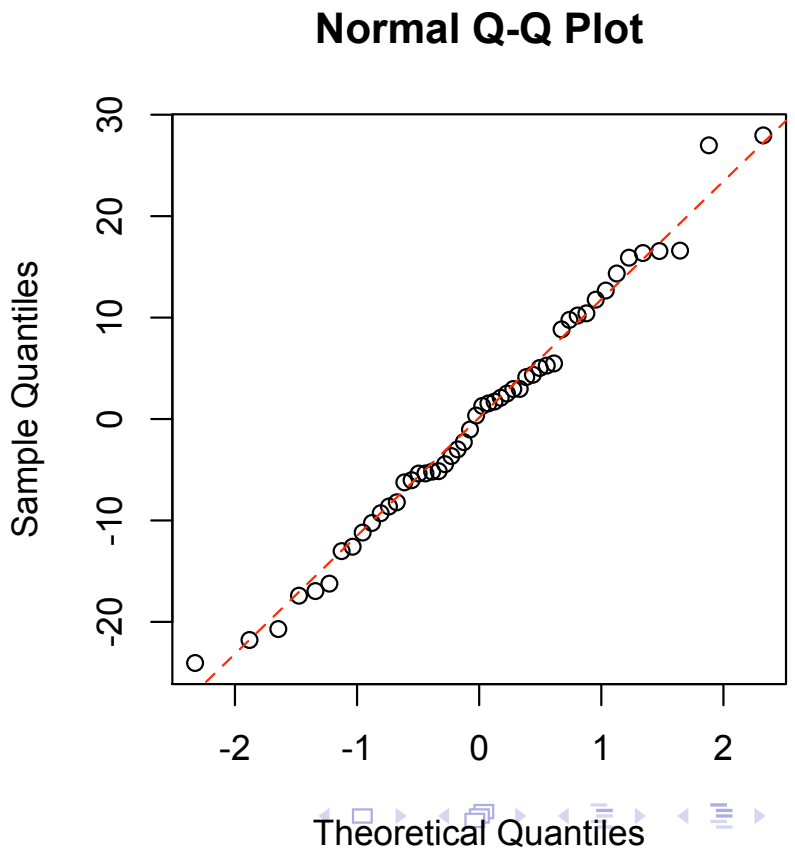
Le q-quantile de la loi de X est x_q tel que

$$x_q = \min\{x : \mathbb{P}(X \leq x) \geq q\}$$

Normalité des résidus?

Le Q-Q Plot permet de comparer les quantiles des résidus avec les quantiles d'une loi gaussienne

```
> qqnorm(reg1$residuals)
```



Inspection des résidus (2bis/4)

Renormalisation

On a $\hat{\xi} = (I - P)\xi$ avec P est la matrice de projection sur l'image de X.

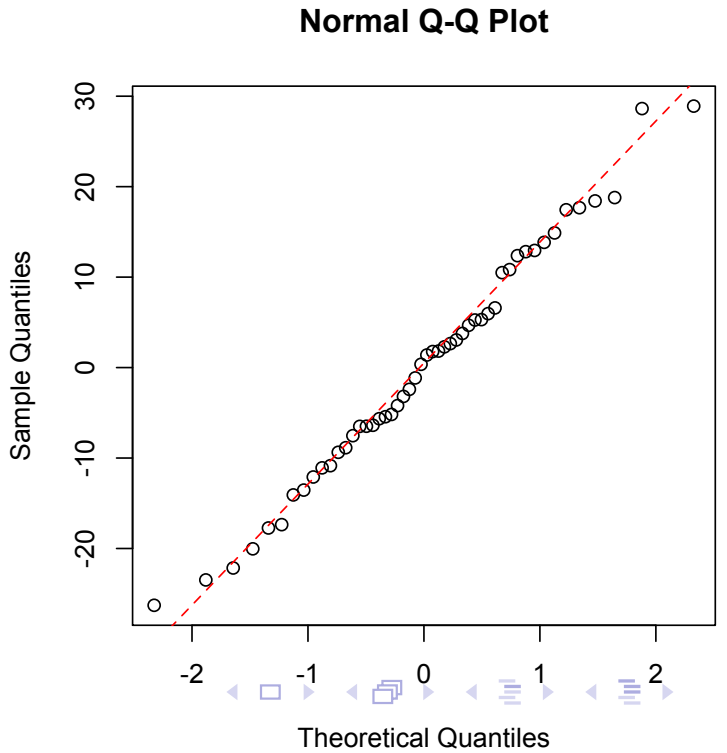
Donc si $\xi_1, \ldots, \xi_n$ i.i.d. de loi $\mathcal{N}(0, \sigma^2)$, les résidus $\hat{\xi}_j$ ont pour loi $\mathcal{N}(0, (1 - P_{jj})\sigma^2)$.

Normalité du bruit?

Il faut regarder le Q-Q Plot des résidus renormalisés

$$\tilde{\xi}_j = (1 - P_{jj})^{-1/2}\hat{\xi}_j$$

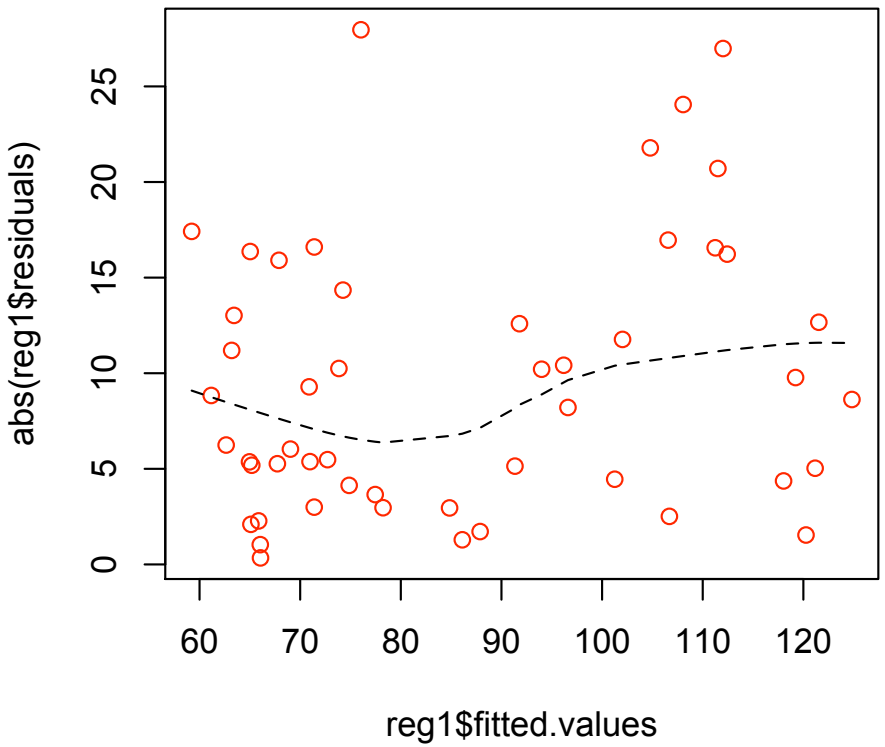
qui ont pour loi $\mathcal{N}(0, \sigma^2)$.



Inspection des résidus (3/4)

La variance dépend-elle du signal?

```
> plot(reg1$fitted.values,abs(reg1$residuals), col=2)
> lines(lowess(reg1$fitted.values,abs(reg1$residual),f=0.7))
```

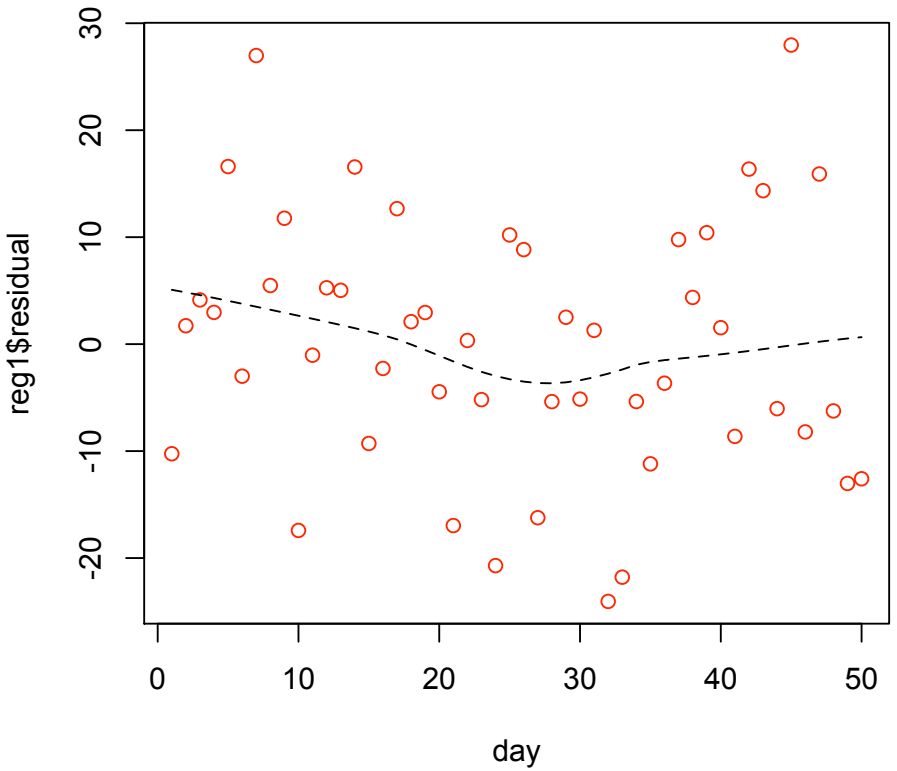


Homoscédasticité des résidus

Inspection des résidus (4/4)

Structuration temporelle des résidus?

```
> plot(reg1$residual,col=2)
> lines(lowess(reg1$residual,f=0.7),lty=2)
```



Petite autocorrélation des résidus

Résultats complets

```
> summary(reg1)
```

	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	54.7278	17.2789	3.17	0.0029	**
T12	-0.3518	1.5731	-0.22	0.8242	
T15	1.4972	1.5377	0.97	0.3361	
Ne12	-4.1922	1.0638	-3.94	0.0003	***
N12	1.2755	1.3632	0.94	0.3551	
S12	3.1711	1.9108	1.66	0.1048	
E12	0.5277	1.9427	0.27	0.7873	
W12	2.4749	2.0720	1.19	0.2393	
Vx	0.6077	0.4858	1.25	0.2182	
O3v	0.2454	0.0965	2.54	0.0150	*

```
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1  1
```

Seules les variables Ne12 et O3v semblent pertinentes !

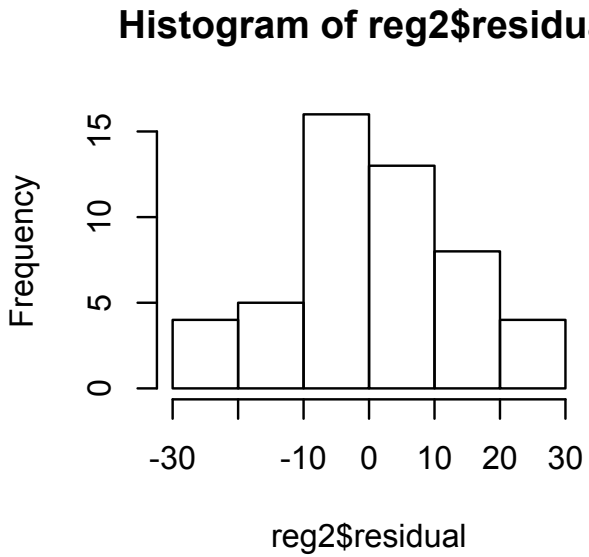
Régression modèle réduit

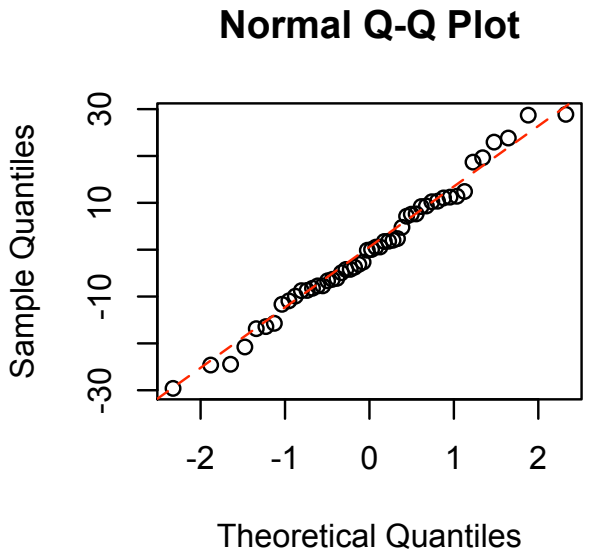
Modèle réduit:
régression de O3 par rapport aux variables Ne12 et O3v

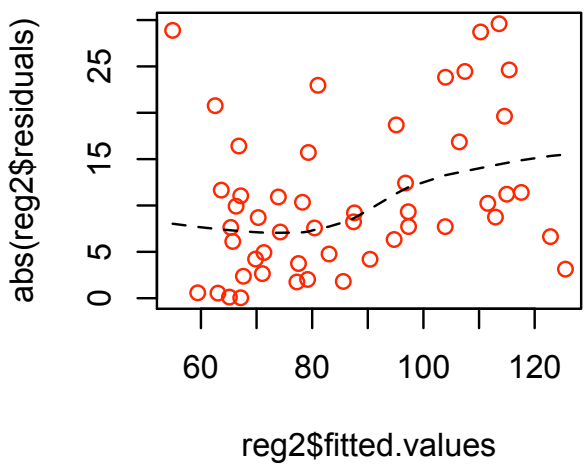
```
> reg2 <- lm(O3~Ne12+O3v,data=ozone)
> summary(reg2)
```

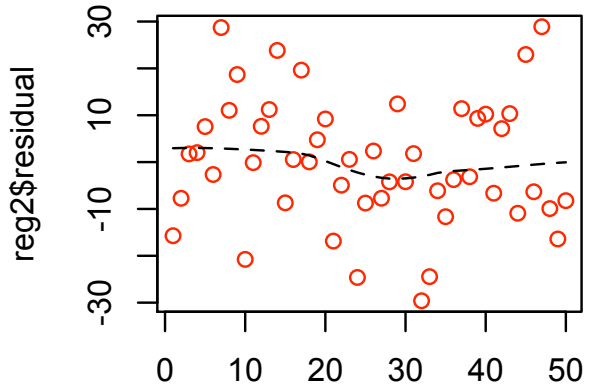
	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	85.0203	11.0943	7.66	0.0000	***
Ne12	-5.4801	0.9102	-6.02	0.0000	***
O3v	0.3416	0.0925	3.69	0.0006	***

Inspection des résidus du modèle réduit









Et quid d'un modèle intermédaire?

Modèle intermédiaire:

```
> reg3 <- lm(O3~T15+Ne12+Vx+O3v,data=ozone)
> summary(reg3)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	61.8252	14.8972	4.15	0.0001	***
T15	1.0577	0.4522	2.34	0.0239	*
Ne12	-3.9935	1.0072	-3.97	0.0003	***
Vx	0.3146	0.1608	1.96	0.0566	.
O3v	0.2629	0.0922	2.85	0.0065	**

Certaines variables déclarées non-importantes dans le modèle complet, sont déclarées importantes dans le modèle intermédiaire...

Quel modèle choisir?

Sélection de modèle

Notations

On associe à chaque sous-ensemble m de variables

- $\widehat{\theta}_m$ obtenu par la régression de O3 par rapport aux variables dans m.
- $\mathcal{L}(m) =$ la log-vraisemblance de $\widehat{\theta}_m$.

Critères classiques

Choix de $\widehat{m}$ vérifiant

$$\widehat{m} \in \underset{m}{\operatorname{argmin}} \{-2\mathcal{L}(m) + \lambda \operatorname{Card}(m)\} \quad \text{avec}$$

critère AIC : $\lambda_{\text{AIC}} = 2$
critère BIC : $\lambda_{\text{BIC}} = \log(n)$

Sélection de modèle : cadre gaussien

Cadre gaussien : $-2\mathcal{L}(m) = n \log(\| Y - X\widehat{\theta}_m \|^2) + \dots$

Critère AIC et BIC : cadre gaussien

$$\widehat{m} \in \operatorname*{argmin}_m \left\{ n \log(\| Y - X\widehat{\theta}_m \|^2) + \lambda \, \mathrm{Card}(m) \right\} \quad \text{avec}$$

AIC : $\lambda_{\mathrm{AIC}} = 2$
BIC : $\lambda_{\mathrm{BIC}} = \log(n)$

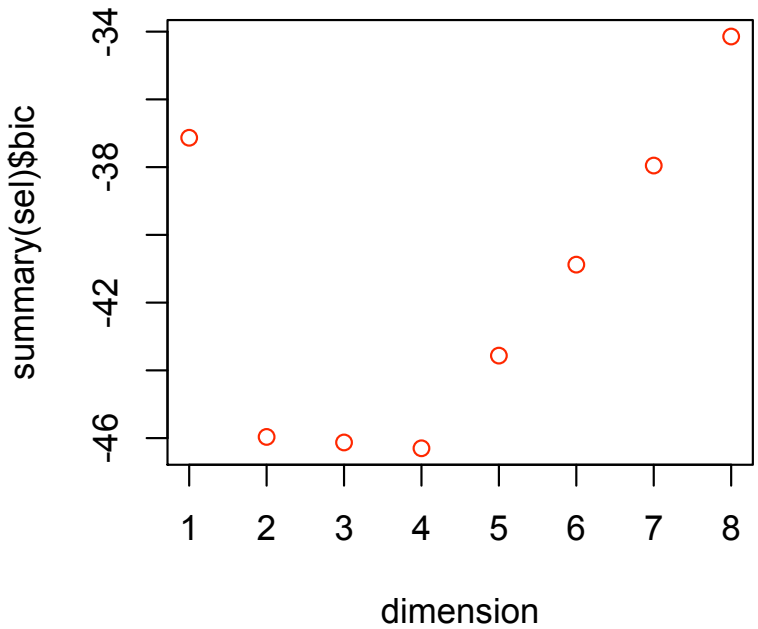
Attention!

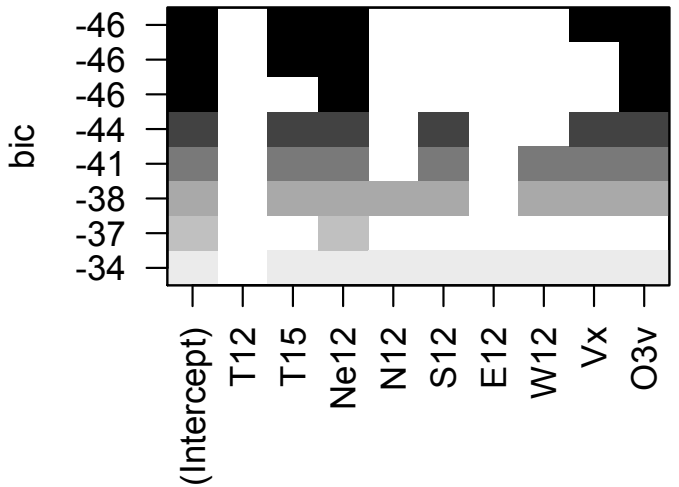
AIC et BIC sont très utilisés, mais leur justification est asymptotique: valables uniquement dans le cas $n \gg k$
où $n =$ taille d'échantillon et $k =$ nombre de variables

Une analyse non-asymptotique donne le choix: $\lambda \approx 1 + 2\log(k)$

Sélection de modèle avec BIC

```
> library(leaps)
> sel<-regsubsets(O3~.,method="exhaustive",data=ozone)
> par(mfrow=c(1,2))
> plot(summary(sel)$bic,xlab="dimension",col=2)
> plot(sel)
```





Régression partielle (1/3)

Questions

- le modèle linéaire par rapport à la variable j est-il raisonnable?
- quelle est l'influence de la variable j?

Régression partielle par rapport à la variable j

Si $Y = \sum_k \theta_k X_k + \xi$ alors

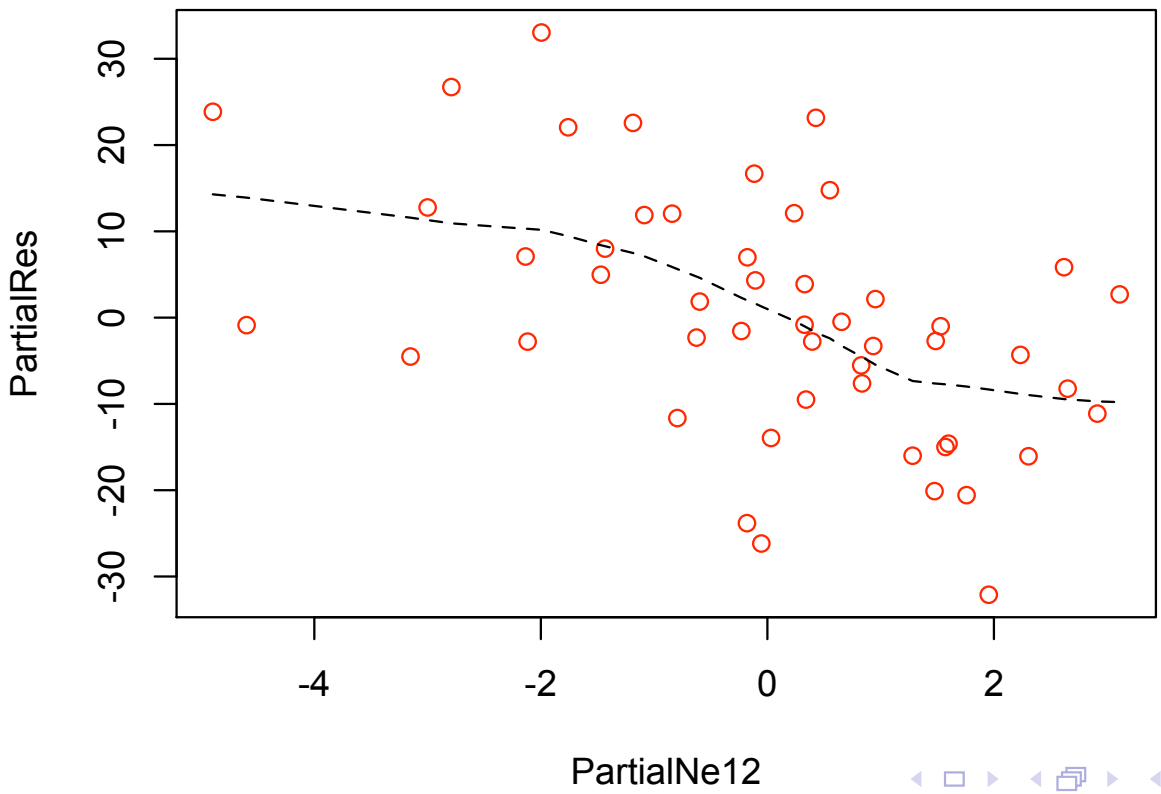
$$\text{lm}(Y \sim -j)\$\text{residuals} = \theta_j \, \text{lm}(j \sim -j)\$\text{residuals} + \text{lm}(\xi \sim -j)\$\text{residuals}$$

Cette relation permet d'inspecter graphiquement la relation de linéarité entre Y et X_j

Régression partielle (2/3)

Régression partielle par rapport à Ne12:

```
> PartialRes <- lm(O3~T15+Vx+O3v,data=ozone)$residuals
> PartialNe12 <- lm(Ne12~T15+Vx+O3v,data=ozone)$residuals
> plot(PartialNe12,PartialRes, col=2)
> lines(lowess(PartialNe12,PartialRes,f=0.7), lty=2)
```



Régression partielle (3/3)

Régression partielle par rapport à T15:

```
> PartialRes <- lm(O3~Ne12+Vx+O3v,data=ozone)$residuals
> PartialT15 <- lm(T15~Ne12+Vx+O3v,data=ozone)$residuals
> plot(PartialT15,PartialRes, col=2)
> lines(lowess(PartialT15,PartialRes,f=0.7), lty=2)
> plot(O3~T15,data=ozone,col=2)
> lines(lowess(ozone[c("T15","O3")],f=0.7), lty=2)
```