Modèle Linéaire Gaussien
Chapitre 4 – Analyse de la covariance

Achille Thin

09 Janvier 2024

Executive Master Statistique et Big Data

Objectifs du Chapitre

Cadre de l'analyse de la covariance (ANCOVA): expliquer les variations d'**une variable quantitative**, appelée variable réponse, **en fonction de variables quantitatives ET qualitatives**, appelées variables explicatives.

Démarche statistique:

1. Écriture du modèle
2. Ajustement (estimation) du modèle grâce aux données
3. Vérification de la validité des hypothèses faites dans le modèle
4. Test de la pertinence des différents éléments du modèle
5. Critique du modèle
6. Conclusion

Objectifs

Question : le nombre d'années d'étude et le genre ont-t-ils une influence sur le salaire horaire (en échelle log) ?

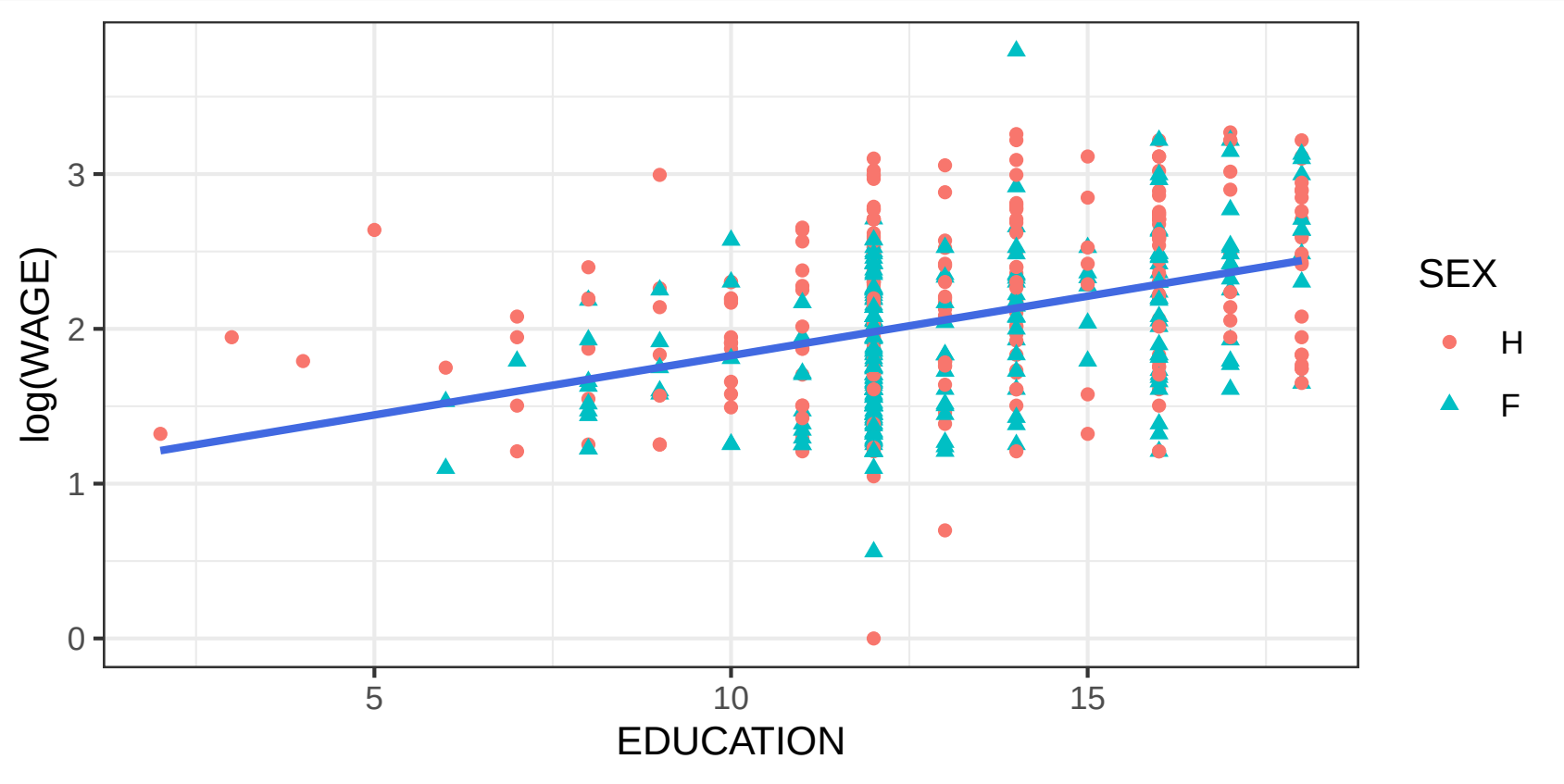
- **Variable réponse:** le salaire horaire en échelle log (log(WAGE)).
- **Variables explicatives:** le nombre d'années d'étude (EDUCATION) et le genre (SEX).

Chargement des données:

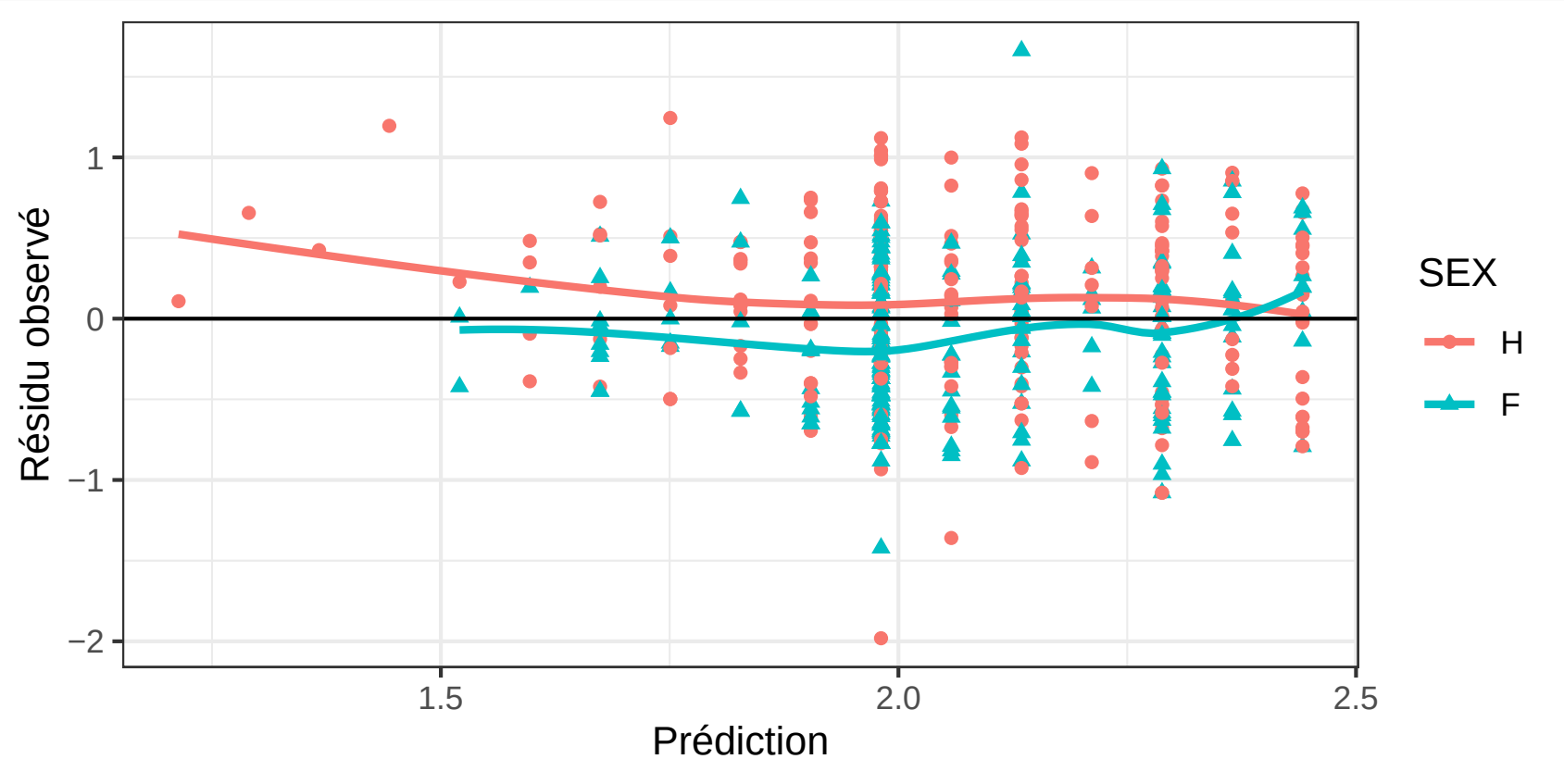
```
donnees <- read.table("../data/wagesmicrodata.csv",
                      sep = ";", dec = ",", header = TRUE,
                      colClasses = c(rep("numeric", 2),
                                     "factor",
                                     rep("numeric" , 3),
                                     rep("factor", 6)))
attach(donnees)
levels(donnees$SEX) <- c(levels(donnees$SEX), "H", "F")
donnees$SEX[donnees$SEX == "0"] <- "H"
donnees$SEX[donnees$SEX == "1"] <- "F"
```

Régression linéaire simple : droite de régression

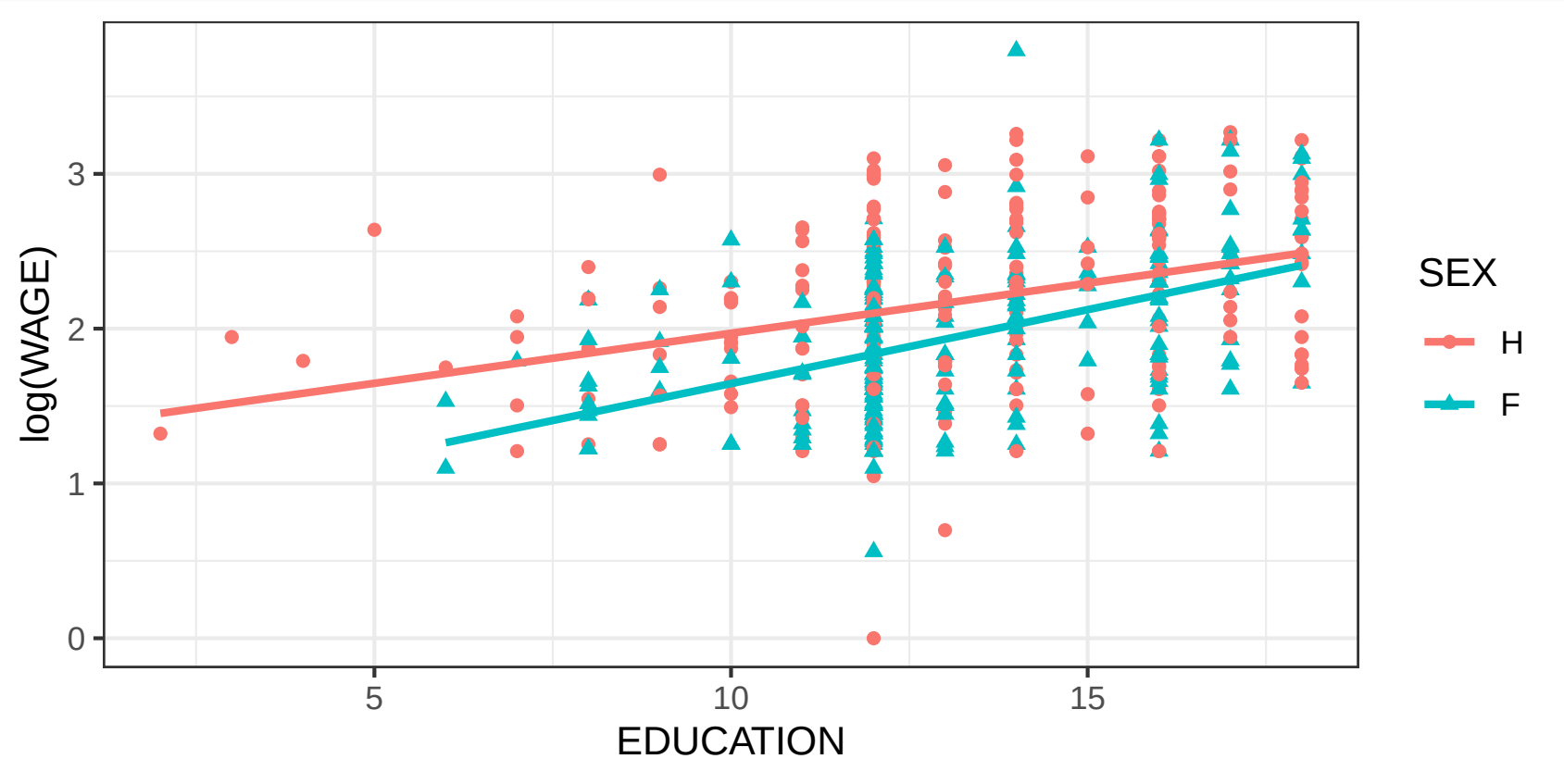
```
reg_simple <- lm(log(WAGE) ~ EDUCATION)
```



Régression linéaire simple : résidus par facteur



Régression linéaire simple : droites de régression par facteur



Écriture du modèle

Notations

Variables explicatives:

- Le facteur a I modalités.
- Chaque modalité facteur est codée par un entier i, $i \in [\![1, I]\!]$.
- Pour la modalité i, on dispose de n_i observations.
- Pour la variable quantitative, on note $x_{i,j}$ la j-ème observation du régresseur pour la modalité i.

Exemple: sur nos données, $I = 2$, $F = 1$, $H = 2$. On a $n_1 = 534$ et $n_2 = 534$. $x_{1,5}$ désigne le nombre d'années d'étude pour la 5-ème femme sondée.

Variable réponse: on note $y_{i,j}$ la valeur de la variable réponse pour la j-ème observation de la modalité i du facteur.

Exemple: sur nos données $y_{1,5}$ désigne le salaire horaire de la 5-ème femme sondée.

Modèle régulier

On suppose que $y_{i,j}$ est la réalisation d'une variable aléatoire $Y_{i,j}$ telle que :

$$Y_{i,j} = b_i + a_i x_{i,j} + \varepsilon_{i,j}, \quad 1 \leq i \leq I = 2, \quad 1 \leq j \leq n_i,$$

- b_i (paramètre inconnu) est l'ordonnée à l'origine de la régression simple pour la modalité i,
- a_i (paramètre inconnu) est la pente de la régression simple pour la modalité i (elle représente l'effet de l'éducation sur le salaire pour la modalité i),
- $\varepsilon_{i,j}$ est une variable aléatoire appelée **bruit**, telle que toutes les variables aléatoires $(\varepsilon_{i,j})$ sont **indépendantes**, d'**espérance nulle** et ont la **même variance**, égale à σ^2 (paramètre inconnu).

Cas particulier du modèle linéaire gaussien: les variables aléatoires $\varepsilon_{i,j}$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Pour chaque modalité i, on obtient les résultats de la régression linéaire simple.

Estimateurs de a_i et b_i (variable aléatoire)

$$\widehat{A}_i = \frac{\sum_{j=1}^{n_i}(x_{i,j} - \overline{x}_{i\bullet})(Y_{i,j} - \overline{Y}_{i\bullet})}{\sum_{j=1}^{n_i}(x_{i,j} - \overline{x}_{i\bullet})^2}, \quad \widehat{B}_i = \overline{Y}_{i\bullet} - \widehat{A}\overline{x}_{i\bullet}.$$

Prédicteur: (variable aléatoire)

$$\widehat{Y}_{i,j} = \widehat{B}_i + \widehat{A}_i x_{i,j}$$

Estimateur de σ^2 (variable aléatoire)

$$S^2 = \frac{1}{n - 2I} \sum_{i=1}^{I} \sum_{j=1}^{n_j} (Y_{i,j} - \widehat{Y}_{i,j})^2.$$

Modèle singulier

Pour des questions d'interprétation, on décompose les paramètres de régression sous la forme $a_i = \beta + \gamma_i$ et $b_i = \mu + \alpha_i$:

- μ et β (paramètres inconnus) : paramètres de la droite de référence,
- α_i (paramètres inconnus) : modification de l'ordonnées à l'origine dûe à la modalité i,
- γ_i (paramètres inconnus) : modification de la pente dûe à la modalité i.

On suppose alors que $y_{i,j}$ est la réalisation d'une variable aléatoire $Y_{i,j}$ telle que :

$$Y_{i,j,k} = \mu + \alpha_i + (\beta + \gamma_i)x_{i,j} + \varepsilon_{i,j}, \quad 1 \leq i \leq I = 2, \quad 1 \leq j \leq n_i.$$

Cas particulier du modèle linéaire gaussien: les variables aléatoires $\varepsilon_{i,j}$ sont indépendantes et identiquement distribuées de loi $\mathcal{N}(0, \sigma^2)$.

Modèle singulier

Cette nouvelle écriture conduit à un problème d'identifiabilité :

- On a décomposé $2I$ paramètres en $2I + 2$ paramètres. Mais on ne dispose que de $2I$ équations $\rightsquigarrow$ singularité du modèle.
- Ayant 2 paramètres « en trop », il faut imposer 2 contraintes pour pouvoir ajuster le modèle. On utilise la droite de régression du 1er niveau du facteur comme référence, *i.e.*,

$$\alpha_1 = \gamma_1 = 0.$$

C'est la **contrainte par défaut de R !**

Estimateur et estimation

Important: les estimateurs des paramètres inconnus **dépendent des contraintes choisies**. On considère ici la contrainte par défaut de R : $\alpha_1 = \gamma_1 = 0$.

Estimateurs (variables aléatoires). Les estimateurs de μ, α_i, β et γ_i sont donnés par

$$\widehat{\mu} = \widehat{B}_1, \quad \widehat{\alpha}_i = \widehat{B}_i - \widehat{B}_1, \quad \widehat{\beta} = \widehat{A}_1, \quad \widehat{\gamma}_i = \widehat{A}_i - \widehat{A}_1.$$

Estimations (valeurs numériques calculées sur les données)

$$\widehat{\mu}^{\text{obs}} = \widehat{b}_1, \quad \widehat{\alpha}_i^{\text{obs}} = \widehat{b}_i - \widehat{b}_1, \quad \widehat{\beta}^{\text{obs}} = \widehat{a}_1, \quad \widehat{\gamma}_i^{\text{obs}} = \widehat{a}_i - \widehat{a}_1.$$

Avec R:

```
reg <- lm(log(WAGE) ~ SEX * EDUCATION)
reg$coefficients  # ou coef(reg)
```

```
   (Intercept)           SEX1      EDUCATION SEX1:EDUCATION
      1.323516      -0.633150       0.064683       0.030805
```

Variables ajustées, résidus et estimateur de σ^2

Ces quantités restent **inchangées** quelques soient les contraintes choisies.

Variables et valeurs ajustées pour $1 \leq i \leq I$, $1 \leq j \leq n_i$

$$\widehat{Y}_{i,j} = \widehat{\mu} + \widehat{\alpha}_i + (\widehat{\beta} + \widehat{\gamma}_i)x_{i,j} = \widehat{B}_i + \widehat{A}_i x_{i,j} \quad \text{et} \quad \widehat{y}_{i,j} = \widehat{y}_i + \widehat{a}_i x_{i,j}.$$

Résidus et résidus observés pour $1 \leq i \leq I$, $1 \leq j \leq n_i$:

$$\widehat{\varepsilon}_{i,j} = Y_{i,j} - \widehat{Y}_{i,j} \quad \text{et} \quad \widehat{e}_{i,j} = y_{i,j} - \widehat{y}_{i,j}.$$

Estimateur et estimation de la variance du bruit (σ^2)

$$S^2 = \frac{1}{n-2I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{i,j} - \widehat{Y}_{i,j})^2 \quad \text{et} \quad \widehat{\sigma}^2 = \frac{1}{n-2I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (y_{i,j} - \widehat{y}_{i,j})^2.$$

Cet estimateur vérifie

$$(n-2I)\frac{S^2}{\sigma^2} \sim \chi^2(n-2I).$$

Validité des hypothèses

Diagnostique des résidus

Avant d'analyser les sorties du modèle ajusté: il faut regarder si les hypothèses du modèle linéaire gaussien sont vérifiées sur nos données, *i.e.*, les variables aléatoires $\varepsilon_{i,j}$

(P1) sont **indépendantes**,

(P2) sont toutes d'**espérance nulle** ($\rightsquigarrow$ la relation entre $\mathbf{y}$ et $\mathbf{x}$ est bien affine),

(P3) ont la **même variance** σ^2 (homoscédasticité),

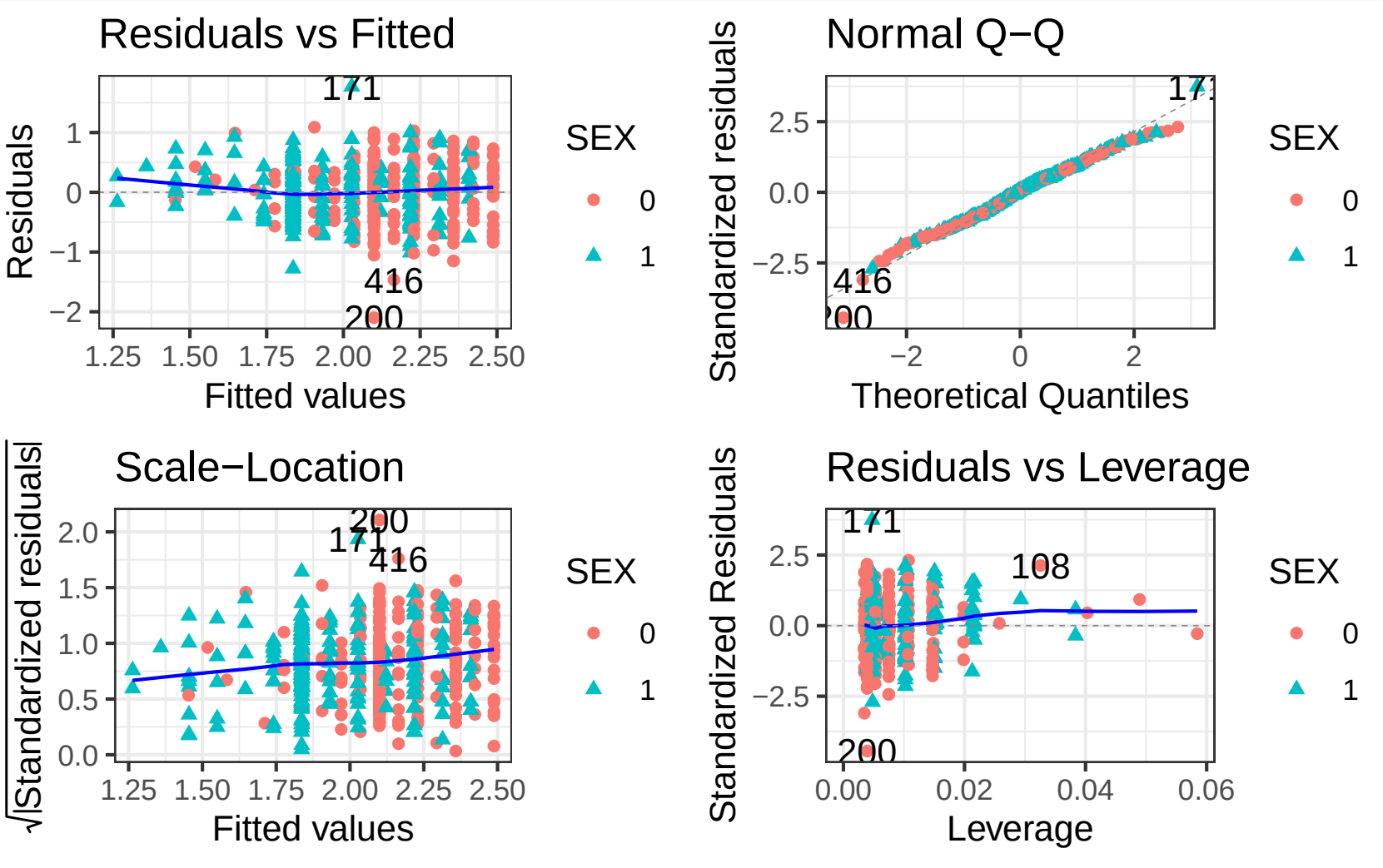
(P4) suivent une **loi normale**.

Validation des hypothèses:

- (P1) : l'indépendance ne peut être assurée que par le protocole expérimentale.
- (P2), (P3), (P4) : on fait la même analyse graphique des résidus observés que pour la régression. Pour (P2) et (P3), on veut le même comportement pour toutes les modalités.

Avec R : 4 graphiques à analyser

```
par(mfrow = c(2, 2))
plot(reg, col = SEX, pch = 16)
```



Tests d'hypothèses

Test du modèle ou test de Fisher global

Question: le nombre d'années d'étude et le genre ont-t-ils une influence sur le salaire horaire (en échelle log) ?

Mathématiquement, cela revient à tester

$$\mathcal{H}_0 : Y_{i,j,k} = \mu + \varepsilon_{i,j,k}, \quad \text{contre} \quad \mathcal{H}_1 : Y_{i,j,k} = \mu + \alpha_i + (\beta + \gamma_i)x_{i,j} + \varepsilon_{i,j}.$$

- Le modèle réduit (celui associé à $\mathcal{H}_0$) ne fait intervenir que l'intercept et le bruit.
- Il s'agit du test de Fisher global vu dans le Chapitre 2 (pour $2I - 1$ paramètres testés). On connaît donc la loi de la statistique de test sous $\mathcal{H}_0$ (loi de Fisher $\mathcal{F}(2I-1, n-2I)$) ainsi que la zone de rejet du test pour un niveau β !

Avec R : test du modèle

```
summary(reg)
```

```
Call:
lm(formula = log(WAGE) ~ EDUCATION * SEX)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0997 -0.3631  0.0342  0.3316  1.7683

...

Residual standard error: 0.474 on 530 degrees of freedom
Multiple R-squared:  0.198,Adjusted R-squared:  0.194
F-statistic: 43.7 on 3 and 530 DF,  p-value: <2e-16
```

Conclusion: la p-valeur (dernière ligne) est inférieure à 5%. On rejette $\mathcal{H}_0$ au niveau 5%. Le modèle ANCOVA explique mieux les données qu'un modèle avec une concentration constante.

Tests des effets

Question: Chacun des effets dans le modèle est-il indispensable ?

Test de l'effet d'interaction: on teste si la covariable a un effet sur la droite de régression (effet de γ_i).

Test des effets principaux: on peut réaliser deux types de test.

- **Test de type I:** l'ajout de l'effet principal du facteur est-il intéressant par rapport à un modèle constant (ou modèle nul) ?
- **Test de type II:** L'ajout de l'effet principal d'un facteur est-il intéressant par rapport à un modèle comprenant déjà la covariable ?

Une collection de modèle

Pour identifier l'influence de chacun des effets, on va mettre en compétition différents modèles.

$$\mathcal{M}_{\mu} : \quad Y_{i,j} = \mu + \varepsilon_{i,j}, \tag{A}$$

$$\mathcal{M}_{\mu,\alpha} : \quad Y_{i,j} = \mu + \alpha_i + \varepsilon_{i,j}, \tag{B}$$

$$\mathcal{M}_{\mu,\beta} : \quad Y_{i,j} = \mu + \beta x_{i,j} + \varepsilon_{i,j}, \tag{C}$$

$$\mathcal{M}_{\mu,\beta,\gamma} : \quad Y_{i,j} = \mu + (\beta + \gamma_i) x_{i,j} + \varepsilon_{i,j}, \tag{D}$$

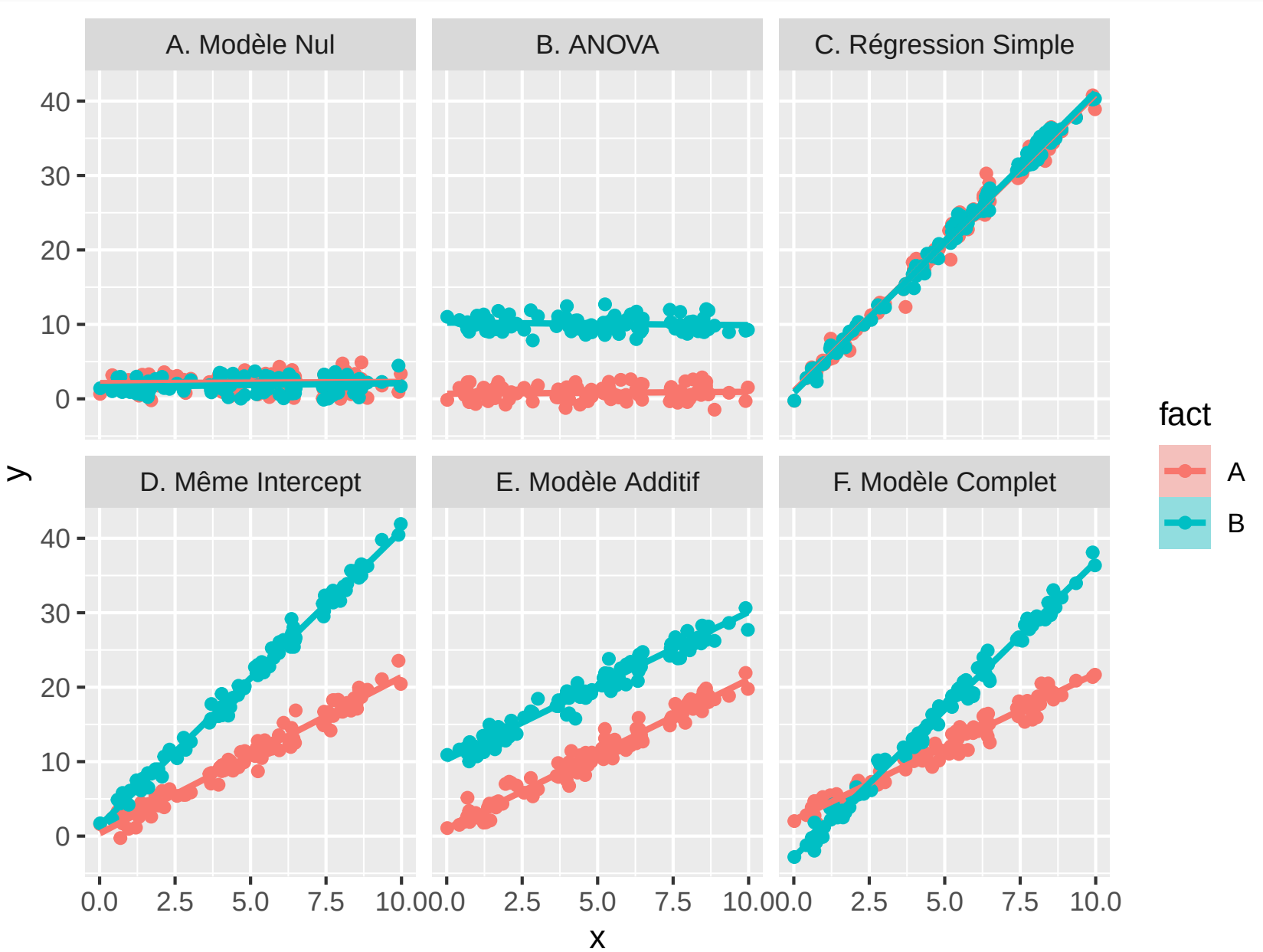
$$\mathcal{M}_{\mu,\alpha,\beta} : \quad Y_{i,j} = \mu + \alpha_i + \beta x_{i,j} + \varepsilon_{i,j}, \tag{E}$$

$$\mathcal{M}_{\mu,\alpha,\beta,\gamma} : \quad Y_{i,j} = \mu + \alpha_i + (\beta + \gamma_i) x_{i,j} + \varepsilon_{i,j}. \tag{F}$$

Remarques:

- $\mathcal{M}_{\mu}$: modèle nul.
- $\mathcal{M}_{\mu,\alpha}$: ANOVA à 1 facteur (la covariable n'a pas d'effet).
- $\mathcal{M}_{\mu,\beta}$: régression linéaire simple (le facteur n'a pas d'effet).
- $\mathcal{M}_{\mu,\alpha,\beta}$: modèle additif.
- $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$: modèle complet.

Une collection de modèle : différents comportement en image



Test pour des modèles emboîtés

Modèles emboîtés: on dit qu'un modèle M_0 est emboîté dans un modèle M_1, lorsque M_1 s'obtient en ajoutant des paramètres à M_0. On note $M_0 \subset M_1$.

Test de Fisher: pour tester l'intérêt d'un modèle M_1 par rapport à un modèle M_0 tel que $M_0 \subset M_1$, on considère

$$\mathcal{H}_0 : \text{le vrai modèle est } M_0 \quad \text{contre} \quad \mathcal{H}_1 : \text{le vrai modèle est } M_1.$$

Statistique de test et zone de rejet au niveau β:

$$F = \frac{(SCR_0 - SCR_1)/q}{\underbrace{SCR_1/(n-p)}_{\text{Estimateur de } \sigma^2 \text{ pour } M_1}} \overset{\mathcal{H}_0}{\sim} \mathcal{F}(q, n-p), \quad \text{et} \quad \mathcal{R} = \left\{ F > q_{1-\beta}^{\mathcal{F}(q,n-p)}, \right\}$$

- q est le nombre de paramètres supplémentaires à estimer dans M_1 par rapport à M_0 et p est le nombre de paramètres à estimer dans M_1,
- SCR_0 et SCR_1 sont respectivement la somme des carrés résiduelles pour M_0 et M_1.

p-valeur: $\mathbb{P}[F > f_{obs}]$, où où f_{obs} est la valeur observée de F.

Avec R : modèle additif ou non ?

Test réalisé:

$\mathcal{H}_0$: le vrai modèle est $\mathcal{M}_{\mu,\alpha,\beta}$ contre $\mathcal{H}_1$: le vrai modèle est $\mathcal{M}_{\mu,\alpha,\beta,\gamma}$.

Modèle $\mathcal{M}_{\mu,\alpha,\beta}$:

```
reg_add <- lm(log(WAGE) ~ SEX + EDUCATION)
```

Lecture de la table:

Effet	Res.Df	RSS	df	Sum of Sq	F	Pr(>F)
$\mathcal{M}_{\mu,\alpha,\beta}$	n - (p - q)	SCR_0				
$\mathcal{M}_{\mu,\alpha,\beta,\gamma}$	$n-p$	SCR_1	q	$SCR_0 - SCR_1$	f_{obs}	p-valeur = $\mathbb{P}[F > f_{obs}]$

Avec la contrainte par défaut de R, $q = I - 1$ et $p = 2I$.

Avec R : test de l'effet d'interaction

Résultat du test:

```
anova(reg_add, reg)
```

```
Analysis of Variance Table

Model 1: log(WAGE) ~ SEX + EDUCATION
Model 2: log(WAGE) ~ SEX * EDUCATION
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1    531 120
2    530 119  1     0.827 3.68  0.056 .
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On conserve l'hypothèse nulle au niveau 5% (p-valeur $>$ 5%). Il n'y a pas d'effet d'interaction lié au facteur.

Que se passe-t-il avec la commande suivante ?

```
anova(reg_add)
```

On teste de façon séquentielle $\mathcal{M}_\mu$ v.s. $\mathcal{M}_{\mu,\alpha}$ et $\mathcal{M}_{\mu,\alpha}$ v.s. $\mathcal{M}_{\mu,\alpha,\beta}$.

Lecture de la table:

Effet	Df	Sum Sq	Mean Sq	F value	Pr(>F)
$\mathcal{M}_{\mu,\alpha}$	I - 1	$SCR_\mu - SCR_{\mu,\alpha}$	Sum Sq/Df	f_{obs}	$\mathbb{P}[F > f_{obs}]$
$\mathcal{M}_{\mu,\alpha,\beta}$	J - 1	$SCR_{\mu,\alpha} - SCR_{\mu,\alpha,\beta}$	Sum Sq/Df	f_{obs}	$\mathbb{P}[F > f_{obs}]$
Residuals	$n - p$	$SCR_{\mu,\alpha,\beta}$	$\widehat{\sigma}^2$		

Avec la contrainte par défaut de R, on a $p = 2I - 1$.

Avec R : anova et tests séquentiels

```
anova(reg_add)
```

```
Analysis of Variance Table

Response: log(WAGE)
           Df Sum Sq Mean Sq F value  Pr(>F)
SEX         1    7.1    7.09    31.4 3.3e-08 ***
EDUCATION   1   21.5   21.53    95.4 < 2e-16 ***
Residuals 531  119.8    0.23
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: l'ajout d'un effet principal pour le genre est pertinent. L'ajout d'un effet principal pour le nombre d'années d'études à un modèle factoriel avec le genre est pertinent.

Que se passe-t-il avec la commande suivante ?

```
library(car)
Anova(reg)
```

On teste

$$\mathcal{M}_{\mu,\alpha} \text{ v.s. } \mathcal{M}_{\mu,\alpha,\beta}, \quad \mathcal{M}_{\mu,\beta} \text{ v.s. } \mathcal{M}_{\mu,\alpha,\beta} \quad \text{et} \quad \mathcal{M}_{\mu,\alpha,\beta} \text{ v.s. } \mathcal{M}_{\mu,\alpha,\beta,\gamma}.$$

Lecture de la table:

Effet	Sum Sq	Df	F value	Pr(>F)
$\mathcal{M}_{\mu,\alpha,\beta}$	$SCR_{\mu,\beta} - SCR_{\mu,\alpha,\beta}$		f_{obs}	$\mathbb{P}[F > f_{obs}]$
$\mathcal{M}_{\mu,\alpha,\beta}$	$SCR_{\mu,\alpha} - SCR_{\mu,\alpha,\beta}$		f_{obs}	$\mathbb{P}[F > f_{obs}]$
$\mathcal{M}_{\mu,\alpha,\beta,\gamma}$	$SCR_{\mu,\alpha,\beta} - SCR_{\mu,\alpha,\beta,\gamma}$		f_{obs}	$\mathbb{P}[F > f_{obs}]$
Residuals	$SCR_{\mu,\alpha,\beta,\gamma}$			

Avec la contrainte par défaut de R, on a $p = 2I - 1$.

Avec R : Anova et tests séquentiels

```
Anova(reg)

Anova Table (Type II tests)

Response: log(WAGE)
               Sum Sq  Df F value  Pr(>F)
SEX              7.1   1   31.80 2.8e-08 ***
EDUCATION       21.5   1   95.90 < 2e-16 ***
SEX:EDUCATION    0.8   1    3.68   0.056 .
Residuals      119.0 530
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: tableau synthétise les conclusions suivantes :

- le modèle contenant l'influence conjointe du nombres d'années d'étude et du genre est pertinent,
- il n'y a pas d'effet d'interaction.

Conclusion

Ce dont on n'a pas parlé...

- **Coefficient de détermination:** l'interprétation du R^2 et du R^2 ajusté reste inchangée. Sur notre exemple la capacité prédictive du modèle est très faible
- **Prédicteur et prédiction:** pour chaque modalité, on obtient les mêmes résultats que pour la régression.

Ma feuille de route pour l'analyse de la covariance

1. Charger les données.
2. Exploration des données et calcul de statistiques descriptives.
3. Écrire le modèle linéaire. Appliquer la fonction `lm` aux données pour ajuster le modèle.
4. Analyser les graphes de résidus pour valider ou invalider les hypothèses du modèles.
5. Faire le test du modèle global : si on ne rejette pas l'hypothèse $\mathcal{H}_0$, on arrête là, le modèle linéaire n'est pas adapté.
6. Faire les test des différents effets.
7. Critiquer le modèle, conclure.