COMPLEMENT

1) Vérifions via un test statistique que le coefficient β associé à la vitesse des rafales ne soit pas inférieur à 0,02

$$H_0 : \hat{B}_r = 0,02$$

$$H_1 \quad \hat{B}_r < 0,02$$

$$T = \frac{\hat{B}_r - 0,02}{\hat{\sigma}_{B_r}} \sim t(n-p)$$

On $n > 30$

$$\Rightarrow T = \frac{\hat{B}_r - 0,02}{\hat{\sigma}_{B_r}} \sim N(0,1)$$

$$\text{AN! } t = \frac{0,0214088 - 0,02}{0,00262332}$$

$$= 0,5370$$

$$\text{P\_value} = P(B_r < 0,537 \mid H_0) = 0,7$$

La p-value est élevée donc on rejette H_0

2) Faisons le test sur l'ecart-type residuel

$H_0 : \sigma = 0{,}5$

$H_1 \quad \sigma > 0{,}5$

$\alpha = 5\%$

$$T = \frac{SCR}{\sigma^2} \sim \chi(n)$$

Cherchons SCR

$$\hat{\sigma}^2 = \frac{SCR}{n-p}$$

$$\Rightarrow SCR = (n-p)\,\hat{\sigma}^2$$

AN : $SCR = 440 \times (0{,}41810)^2$

$$\underline{\underline{= 76{,}91}}$$

$$t = \frac{76{,}91}{(0{,}5)^2} = 307{,}66$$

$$Ic = [\,k, +\infty\,[$$

$$P(T > k) = \alpha$$

$$P(T \leq k) = 1-\alpha$$

$$P(T < k) = 0{,}95$$

Lecture de table $k = 489{,}9$

$I_c = [489{,}7 , +\infty[$

$t = 367{,}66 \notin I_c$

Donc on accepte H_0

3) Démontrons théoriquement que la vraisemblance du modèle complet est toujours supérieure à celle du modèle sélectionné emboîté

Soient deux modèles M et $M+1$

Considérons que M correspond à un modèle avec $|M|$ variables et $M+1$ correspond au modèle M auquel on a rajouté une variable supplémentaire

Nous avons :

$$R^2(M) = \frac{\| \hat{Y}(M) - \bar{y}\mathbb{1} \|^2}{\| Y - \bar{y}\mathbb{1} \|^2} = 1 - \frac{SCR(M)}{SCT}$$

$$R^2(M+1) - R^2(M) = \frac{\| P_{X_M^\perp} Y \|^2 - \| P_{X_{M+1}^\perp} Y \|^2}{\| Y - \bar{y}\mathbb{1} \|^2}$$

$$= \frac{\| P_{X_M}^{\perp} P_{X_{M+1}}^{\perp} Y + P_{X_M}^{\perp} P_{X_{M+1}}^{\perp} Y \|^2 - \| P_{X_{M+1}} Y \|^2}{\| Y - \bar{y} \mathbb{1} \|^2}$$

$$= \frac{\| P_{X_M}^{\perp} P_{X_{M+1}} Y \|^2}{\| Y - \bar{y} \mathbb{1} \|^2} \geq 0$$

Donc $R^2(M+1) > R^2(M)$

Le R^2 augmente avec le nombre de variable

On sait que :

$$SCR = SCT - SCE$$

$$\Rightarrow \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

$$\Rightarrow \frac{SCR}{SCT} = 1 - R^2$$

$$\Rightarrow SCR = (1 - R^2) SCT$$

Lorsque le R^2 augmente le SCR diminue.

Donc le SCR diminue avec l'augmentation du nombre de variables

Cherchons la vraisemblance

$$\mathcal{L}(Y, B, \sigma^2) = \prod_{i=1}^{n} f_Y(y_i)$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} B_j x_j\right)^2\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\|Y - X\beta\|^2\right]$$

$\mathcal{L}(Y, B, \sigma^2)$ maximal pour β estimateur des MC et $\hat{\sigma}^2 = \dfrac{\|Y - X\hat{B}\|^2}{n}$

Nous avons :

$$\max_{\beta, \sigma^2} \mathcal{L}(Y, B, \sigma^2) = \left(\frac{n}{2\pi\|Y - X\hat{B}\|^2}\right) \exp\left(-\frac{n}{2}\right)$$

$$= \frac{n}{2\pi SCR} \exp\left(-\frac{n}{2}\right)$$

$$= \mathcal{L}(Y, \hat{B}, \hat{\sigma}^2)$$

car $\|Y - X\hat{B}\|^2 = SCR$

Sous l'hypothèse nulle, nous avons :

$$\max_{B, \sigma^2} \mathcal{L}_0(Y, B_0, \sigma^2) = \left(\frac{n}{2\pi SCR_0}\right)^{\frac{n}{2}} \exp\left(\frac{-n}{2}\right) \quad \mathrm{II}$$

$$= \mathcal{L}_0(Y, \hat{B}_0, \hat{\sigma}_0^2)$$

On a démontré que :

$R^2(M_{+1}) > R^2(M)$

$SCR(M_{+1}) < SCR(M)$

Donc en considérant les résultats I et II on peut affirmer que la vraisemblance du modèle complet est toujours supérieure à celle du modèle sélectionné emboîté.

4) Faisons le test de vraisemblance entre ces deux modèles.

Nous allons faire un test pour le modèle complet avec toutes les variables y compris l'interaction entre les variables qualitatives et le modèle emboîté sélectionné obtenu à l'aide de la méthode de sélection de variables bacward.

$$H_0 : \beta_1 = \beta_2 = \beta \dots = \beta_q = 0$$

$$H_1 \quad \exists \quad i \in [\![1, q]\!] \mid \beta_i \neq 0$$

$$\Lambda_n = \frac{\max_{\sigma, \beta} \mathcal{L}_0 (Y, \beta_0, \sigma^2)}{\max_{\sigma, \beta} \mathcal{L} (Y, \beta, \sigma^2)}$$

$$-2 \ln \Lambda_n \sim \chi (2)$$

Région critique

$$D_\alpha = \{ Y \in \mathbb{R}^n : \Lambda_n < \lambda_0 \}$$

Cherchons λ_0

comme la région de rejet $\lambda < k$ est équivalent à $-2\ln(\lambda) > k'$

On rejetera à un niveau approximatif α si

$$-2\ln(\lambda) > \chi^2_{1-\alpha}(p-q)$$

$$\Leftrightarrow -2\ln(\lambda) > \chi^2_{95}(12) = 21,02607$$

$$\lambda = \frac{L_0(Y, \hat{B}_0, \hat{\sigma}^2)}{L(Y, \hat{B}, \sigma^2)}$$

$$I_C = [21,02607, +\infty[$$

AN : $\lambda = 0,146$

$$-2\ln(\lambda) = -2\ln(0,146)$$

$$= 8,445$$

$$-2\ln(\lambda) = 8,445 \notin I_C$$

Donc on accepte H_0

Il faut considérer le modèle emboîté sélectionné. Les variables non prises n'ont pas un grand intérêt à priori.

6) Sur la base des 20 premiers résidus du modèle sélectionné, faisons un test d'adéquation à la loi normale centrée et variance qu'on précisera.

Faisons le test d'adéquation à la loi normale $\mathcal{N}(0, \sigma^2)$

avec $\sigma^2 = 0{,}22$ la variance empirique corrigée de l'échantillon.

Les résidus sont continus donc un test de Khi deux serait pas très approprié. On décide de faire le test d'adéquation de Kolmogorov-Smirnov.

$$D = \max_{1 \leq i \leq n} \left\{ \max \left\{ \left| F(X_i) - \frac{i}{n} \right| , \left| F(X_i) - \frac{i-1}{n} \right| \right\} \right\}$$

Voici le tableau des donnés à 10^{-2} près

X_i	$F(X_i)$	$\left\|F(X_i) - \frac{i}{n}\right\|$	$\left\|F(X_i) - \frac{i-1}{n}\right\|$
0,54	0,70	0,65	0,70
0,46	0,68	0,58	0,63
0,32	0,62	0,47	0,52
0,21	0,58	0,38	0,43
0,17	0,57	0,32	0,37
0,13	0,55	0,25	0,30
0,10	0,54	0,19	0,24
0,08	0,53	0,13	0,18
-0,01	0,49	0,04	0,09
-0,04	0,48	0,018	0,03
-0,09	0,46	0,08	0,03
-0,13	0,44	0,15	0,10
-0,14	0,44	0,2	0,15
-0,19	0,42	0,2	0,22
-0,25	0,39	0,35	0,30
-0,33	0,37	0,429	0,37
-0,45	0,32	0,52	0,47
-0,57	0,28	0,62	0,56
-0,88	0,18	0,76	0,71
-1,45	0,06	0,93	0,88

AN: $D = \max_{1 \le i \le n} \left\{ \max \left[\left| F(X_i) - \frac{i}{n} \right| , \left| F(X_i) - \frac{i-1}{n} \right| \right] \right\}$

$= 0{,}93$

$\alpha = 0{,}05$

$D\alpha = 0{,}29408$

$D = 0{,}93 > D\alpha$ donc on rejette

H_0.

NB: Un histogramme est fait en annexe du rapport pour illustrer la forme des 20 premiers residus.

Nous avons un echantillon très faible pour avoir une convergence des residus.