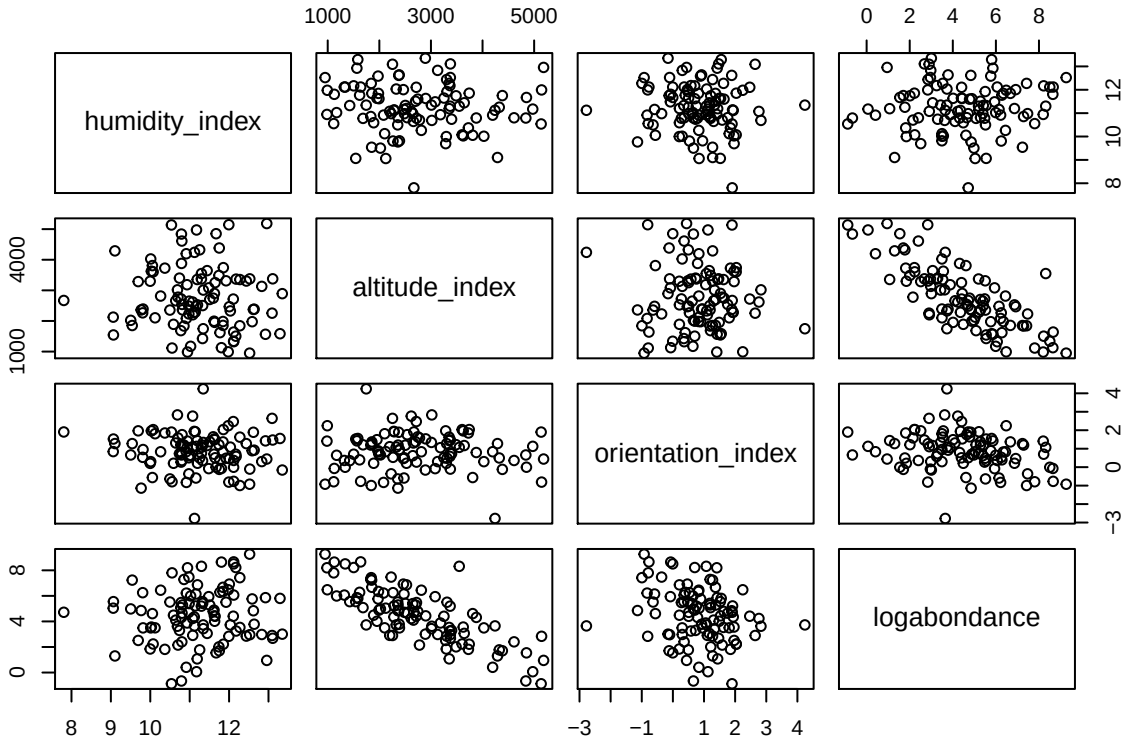


```
lm_estim <- lm(logabondance~ humidity_index + altitude_index + orientation_index,data=data_butterflies)
```

1. [2] Ecrire le modèle statistique correspondant aux commandes R précédentes. Rappeler ses hypothèses.

   - $Y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$ [1]
   - Y_i = logabundance
   - x^1 =humidity index
   - x^2 = altitude index
   - x^3 = orientation index
   - $\varepsilon_i \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$ [1]

2. [3] Commentez les graphes des résidus donnés dans la Figure 1. Les 4 postulats sont ils vérifiés? Que pouvez-vous dire de l'observation $i = 3$?

   - [0.5] En haut à gauche : pas de tendance dans la moyenne (trait rouge horizontal : $\mathbb{E}[\varepsilon_i] = 0$
   - [0.5] En bas à gauche : pas de tendance dans ce graphe :$\mathbb{V}[\varepsilon_i] = cste$
   - [0.5] En haut à droite: points alignés sur bissectrice : résidus gaussiens (sauf $i = 3$!!)
   - [0.5] En bas à droite : pas de point en dehors de la distance de cook. Pas de point abérrant
   - Mais $i = 3$ a un grand résidu standardisé (mal ajusté) [0.5]. Mais petit leverage donc n'influence pas beaucoup l'inférence [0.5]

   Les résultats de l'estimation sont données ci-dessous.

```
summary(lm_estim)
```

```
##
```

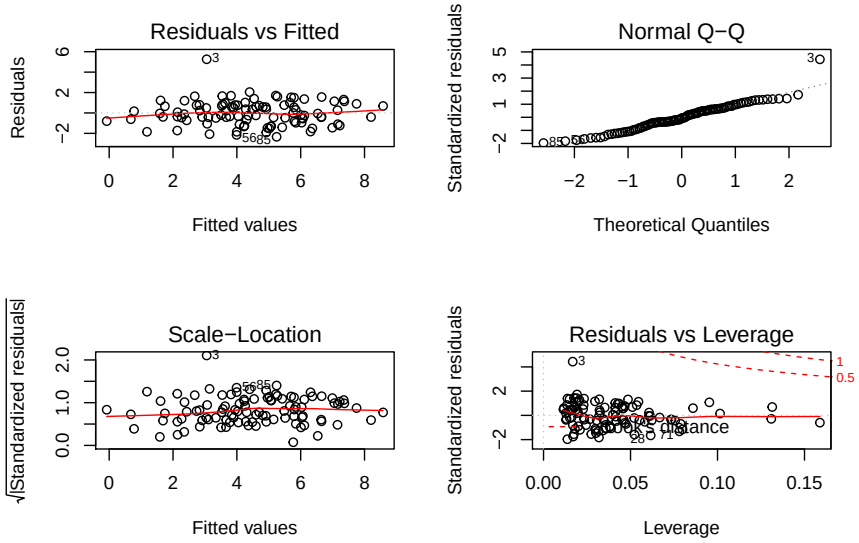

Figure 1: Graphes de résidus

```
## Call:
## lm(formula = logabondance ~ humidity_index + altitude_index +
##     orientation_index, data = data_butterflies)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3368 -0.7446 -0.0876  0.7586  5.2587
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9.4345886  1.4081957   6.700 1.42e-09 ***
## humidity_index    0.0106327  0.1182602   0.090    0.929
## altitude_index   -0.0016437  0.0001171 -14.034  < 2e-16 ***
## orientation_index -0.6217292  0.1196374  -5.197 1.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.197 on 96 degrees of freedom
## Multiple R-squared:  0.6956, Adjusted R-squared:  0.6861
## F-statistic: 73.11 on 3 and 96 DF,  p-value: < 2.2e-16
```

3. 2.5 On cherche à tester le modèle. Rappelez la définition de ce test. Rappelez l'expression de la statistique de test ainsi que sa loi sous l'hypothèse $\mathcal{H}_0$.. Retrouvez les degrés de libertés donnés dans les sorties R précédentes. Donner la conclusion du test du modèle.

- 0.5 $\mathcal{H}_0 : y_i = \beta_0 + \varepsilon_i$ versus $\mathcal{H}_1 : y_i = x_i\beta + \varepsilon_i$
- 0.5 $F = \frac{(\text{SCR}_0 - \text{SCR})/(4-1)}{\text{SCR}/(n-4)}$ où SCR est la somme des carrés résiduels
- 0.5 $F \sim \mathcal{F}\text{isher}(3, n-4)$
- 0.5 $n = 100$, on retrouve 3 et $100 - 4 = 96$
- 0.5 $p - value < 2.2e - 16$, on rejette largement $\mathcal{H}_0$ au niveau 5%.

4. [1] Interpréter la quantité `Pr(>|t|)` pour la covariable `humidity index`.

C'est la p-value du test $\beta_1 = 0$ versus $\beta_1 \neq 0$. Cette p-value $> 5\%$ donc on ne rejette pas $\mathcal{H}_0$ au niveau 5%. Les données ne permettent pas rejetter l'hypothèse que $\beta_1 = 0$ au niveau 5%.

5. [0.5] Donnez l'estimation de σ^2 pour le modèle contenant toutes les covariables.

$\hat{\sigma}^2 = 1.197^2$

6. [5] On souhaite construire un intervalle de confiance pour la quantité $\hat{y}_3 = x_3\beta$ où x_3 sont les covariables de la parcelle 3 et $\hat{\beta}$ est l'estimateur des moindres carrés.

   (a) [1] Rappeler la loi de $\hat{\boldsymbol{\beta}}$ dans le cas du modèle linéaire gaussien.

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

   (b) [1] En déduire la loi de $x_3\hat{\beta}$.

   $x_3\hat{\boldsymbol{\beta}} \sim \mathcal{N}(x_3\boldsymbol{\beta}, x_3\sigma^2(X'X)^{-1}x_3')$

   (c) [2] Construire alors un intervalle de confiance pour $\hat{y}_3$.

$$\frac{x_3\hat{\boldsymbol{\beta}} - x_3\boldsymbol{\beta}}{\sqrt{\sigma^2 x_3(X'X)^{-1}x_3'}} \sim \mathcal{N}(0,1)$$

Or σ^2 est inconnu donc on doit le remplacer par son estimateur.

$$\begin{aligned} T &= \frac{x_3\hat{\boldsymbol{\beta}} - x_3\boldsymbol{\beta}}{\sqrt{\hat{\sigma}^2 x_3(X'X)^{-1}x_3'}} = \frac{\frac{x_3\hat{\boldsymbol{\beta}} - x_3\boldsymbol{\beta}}{\sqrt{\sigma^2 x_3(X'X)^{-1}x_3'}}}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{N}{\sqrt{D}} \quad [0.5] \\ N &= \frac{x_3\hat{\boldsymbol{\beta}} - x_3\boldsymbol{\beta}}{\sqrt{\sigma^2 x_3(X'X)^{-1}x_3'}} \sim \mathcal{N}(0,1) \quad [0.25] \\ D &= \frac{\hat{\sigma}^2}{\sigma^2} \sim \frac{\chi^2(n-4)}{(n-4)} \quad [0.25] \end{aligned}$$

De plus N et D sont indépendantes par le théorème de Cochran [0.25]. Donc $T \sim \mathcal{T}_{n-4}$ [0.25]. Par conséquence, soit $q_{1-\alpha/2,n-4}$ le quantile de niveau $1-\alpha/2$ d'une Student à $n-4$ degrés de liberté. On a

$$\mathbb{P}\left(-q_{1-\alpha/2,n-4} \leq T \leq q_{1-\alpha/2,n-4}\right) = 1-\alpha$$

Donc

$$\mathbb{P}\left(x_3\hat{\boldsymbol{\beta}} - q_{1-\alpha/2,n-4}\sqrt{\hat{\sigma}^2 x_3(X'X)^{-1}x_3'} \leq x_3\boldsymbol{\beta} \leq x_3\hat{\boldsymbol{\beta}} + q_{1-\alpha/2,n-4}\sqrt{\hat{\sigma}^2 x_3(X'X)^{-1}x_3'}\right) = 1-\alpha$$

$$IC = \left[x_3\hat{\boldsymbol{\beta}} - q_{1-\alpha/2,n-4}\sqrt{\hat{\sigma}^2 x_3(X'X)^{-1}x_3'}, x_3\hat{\boldsymbol{\beta}} + q_{1-\alpha/2,n-4}\sqrt{\hat{\sigma}^2 x_3(X'X)^{-1}x_3'}\right] \quad [0.5]$$

   (d) [1] On obtient la fourchettee [2.7473.362] alors qu'on avait observé $y_3 = 8.313263$. Commentez.

On retrouve ici le fait que la prédiction et les valeurs observées sont très différentes. Point qui ne semble pas obéir au modèle mais sans influence notable sur l'inférence.