POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE

DÉPARTEMENT DE MATHÉMATIQUES
ET DE GÉNIE INDUSTRIEL

MTH2302B - PROBABILITÉS ET STATISTIQUE

Devoir - Hiver 2025

Date de remise : 15 avril avant 23h59 (dans Moodle)

Veuillez remplir le tableau suivant et joindre cette page à votre rapport.

Identification de l'étudiant.e 1	
Nom : AKSAS	**Prénom :** Yanis
Groupe : 01	**Matricule :** 2068931

Identification de l'étudiant.e 2	
Nom : PAQUETTE	**Prénom :** Laurie
Groupe : 01	**Matricule :** 2092149

Placer les deux fichiers `DevoirB_H25.csv` et `charger.R` dans le répertoire de travail de `R`. En utilisant votre **matricule**, exécuter ensuite (dans cet ordre) les deux commandes suivantes dans `R` pour générer votre ensemble de données personnalisées 'mondata' :

```
source('charger.R')
mondata <- charger(matricule)
```

Question	Note
a)	/4
b)	/7
c)	/7
d)	/5
e)	/10
f)	/3
g)	/2
Présentation	/2
TOTAL	/40

Mardi le 15 avril 2025

MTH2302B- PROBABILITÉS ET STATISTIQUE

Devoir- Hiver 2025

CONTEXTE

Ce devoir est une étude de cas qui consiste en une analyse des données recueillies au cours d'une expérience visant l'amélioration d'un procédé d'assemblage par rivetage des panneaux du fuselage d'un modèle d'avion.

La coque ainsi que les ailes d'un avion sont constituées de différents panneaux qui tiennent par des rivets. La pose de ces rivets exige le perçage d'un grand nombre de trous dont la

qualité (position, précision, etc.) est primordiale pour l'obtention d'un bon assemblage. Le but de l'étude est de : (1) analyser et évaluer l'importance de l'effet des différentes facteurs variables sur la qualité du perçage; (2) déterminer un modèle mathématique convenable permettant d'évaluer la qualité du perçage en fonction des variables pertinentes; et (3) déterminer (si possible) les valeurs optimales de ces variables pour un perçage de bonne qualité. La description des variables de l'étude est donnée ci-dessous.

La mesure de la qualité du procédé de perçage. La qualité du procédé de perçage est mesurée par celle des trous obtenus. Dans le cadre de cette expérience (voir le Tableau 1), la variable utilisée est l'*Indice de rugosité* (IR) qui est une mesure de la qualité du fini de la surface d'un trou; elle est obtenue à l'aide d'une sonde. Plus l'indice est petit, meilleure est la qualité du perçage.

Les variables susceptibles d'affecter la qualité du perçage. La qualité du procédé de perçage peut dépendre de plusieurs facteurs. Dans cette étude, après un certain nombre d'analyses, quelques variables furent retenues par les ingénieurs responsables du procédé. Chaque trou est obtenu en utilisant une perceuse dont la vitesse de rotation ainsi que l'avance sont contrôlables. Le perçage a lieu sous un jet d'air comprimé dont la température est contrôlable. Ainsi, les quatre variables retenues furent : la *vitesse* de rotation de la perceuse, l'*avance* du foret de la perceuse, la *température* de l'air injecté et le *type de matériau*.

Les données. Les données à analyser ont été obtenues à l'aide d'une expérience planifiée. L'expérience ne pouvant pas être menée sur une véritable ligne d'assemblage, elle fut réalisée avec des spécimens de pièces métalliques appelés coupons. Certains coupons étaient constitués d'un type de matériau (codé 0), et les autres d'un autre type de matériau (codé 1). Les essais furent réalisés en perçant des trous dans les coupons et ce, selon diverses combinaisons de valeurs des facteurs. L'indice (IR) des trous obtenus était mesuré dans chaque cas. Le Tableau 1 ci-dessous présente les différentes variables de l'étude (numéro de colonne dans le fichier, symbole, description, etc.) telles qu'elles apparaissent dans votre ensemble de données personnalisées (constitué de 220 observations sur 5 variables).

Colonne no	Nom (Symbole)	Description
1	Matériau (M)	Le type de matériau du coupon (deux types codés **0** et **1**)
2	Vitesse (V)	La vitesse de la perceuse (en 1000 tours par minute)
3	Avance (A)	L'avance du foret
4	Température (T)	La température (en $^{\circ}$F) de l'air injecté
5	Indice (IR)	L'indice de rugosité (sans unité)

Tableau 1 : Les variables de l'étude.

Le but est d'analyser les résultats issus de cette expérience afin d'évaluer : l'effet de chacun des facteurs sur la qualité des trous obtenus et en retenir les plus influents; la façon par laquelle ces facteurs interagissent et quelle combinaison des valeurs de facteurs donne lieu à un procédé de perçage de qualité optimale.

Importation des données

In [1]:

```
source("data/charger.R")
data <- charger(2068931)
colnames(data)[1] <- "M"
data
```

A data.frame: 220 x 5

	M	V	A	T	IR
	<int>	<int>	<int>	<dbl>	<dbl>
120	0	15	12	40.2	8.2
243	1	10	9	41.1	9.9
146	0	15	12	44.2	14.7
30	1	6	12	47.8	24.9
259	1	6	6	42.1	13.4
237	1	10	9	42.1	12.5
103	0	6	6	42.0	15.0
191	1	15	12	40.5	13.7
8	0	6	9	38.3	6.1
141	0	10	9	39.0	9.0
185	1	6	6	41.2	10.4
188	1	10	9	41.1	8.7
248	1	6	6	40.3	10.0
226	1	15	12	41.4	4.8
104	0	15	12	44.9	10.9
51	0	6	6	35.8	3.8
196	1	6	6	41.6	14.0
261	1	10	9	38.9	8.3
74	0	15	12	38.5	8.4
23	1	6	6	36.8	3.9
168	0	10	9	43.1	10.8
13	0	10	9	40.6	4.8
126	0	10	9	39.4	5.8
90	0	15	12	45.4	13.1
247	1	15	12	40.5	7.8
230	1	15	12	43.5	15.2
204	1	6	6	41.2	10.8
100	0	10	9	42.7	12.6
229	1	10	9	40.2	10.6
213	1	6	6	39.5	6.3
...	...	...	...	...	...
187	1	10	9	38.5	7.7

	M	V	A	T	IR
	<int>	<int>	<int>	<dbl>	<dbl>
236	1	6	6	42.6	15.5
175	1	15	12	39.5	8.2
109	0	15	12	39.3	10.1
266	1	15	12	41.2	10.6
284	1	10	9	43.2	10.1
89	0	10	9	46.0	14.8
9	0	10	6	36.9	3.2
10	0	10	6	38.1	3.4
224	1	6	6	43.5	11.9
177	1	6	6	41.6	8.6
49	0	15	6	36.7	5.5
42	0	15	12	53.4	34.0
156	0	10	9	44.9	16.6
276	1	15	12	44.2	9.7
264	1	15	12	42.5	11.4
154	0	6	6	39.0	10.6
262	1	15	12	39.2	11.1
260	1	10	9	37.3	9.3
167	0	6	6	40.1	12.5
145	0	6	6	40.6	7.5
25	1	15	6	39.4	6.2
114	0	15	12	45.3	17.5
45	0	10	9	37.1	4.9
2	0	6	9	41.3	6.2
124	0	10	9	44.6	15.6
190	1	15	12	41.8	11.9
116	0	6	6	46.5	17.7
122	0	15	12	44.0	18.0
29	1	6	12	48.9	26.0

Phase 1 : Analyse statistique descriptive et inférence

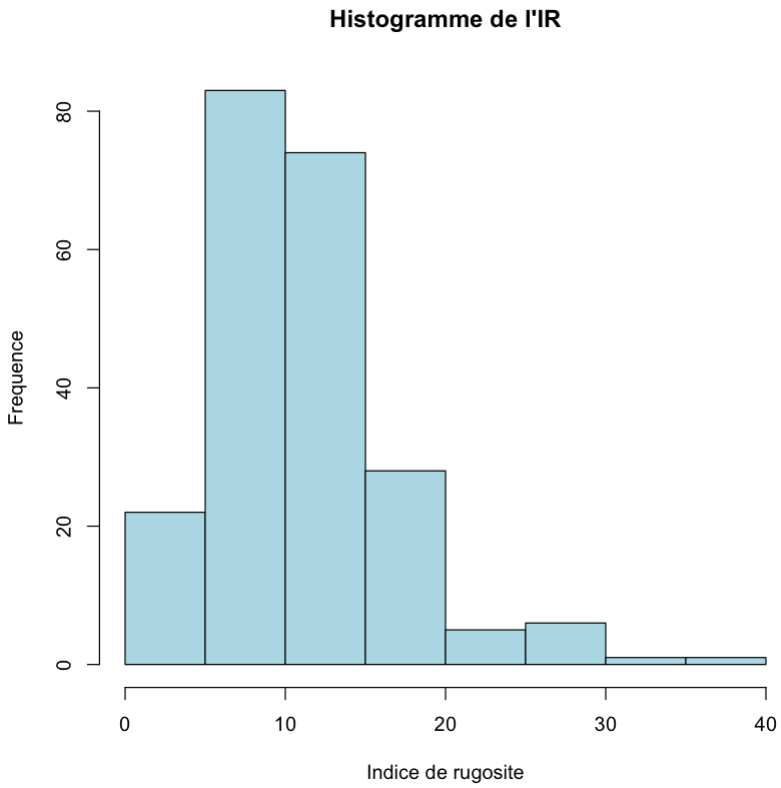
On demande de répondre aux questions suivantes en utilisant des techniques appropriées de statistique (statistique descriptive et inférence), illustrées par des diagrammes pertinents.

a) (4 points) Pour la variable *Indice de rugosité*, produisez les graphiques et les tableaux demandés et interprétez brièvement le résultat dans chaque cas :

- un histogramme et un diagramme de Tukey (ou «Box Plot»);
- une droite de Henry (ou «Normal Probability Plot») et un test de normalité (Shapiro-Wilk);
- un tableau de statistiques descriptives comprenant : *moyenne*, *quartiles*, *écart type*, *intervalle de confiance pour la moyenne*.

HISTOGRAMME

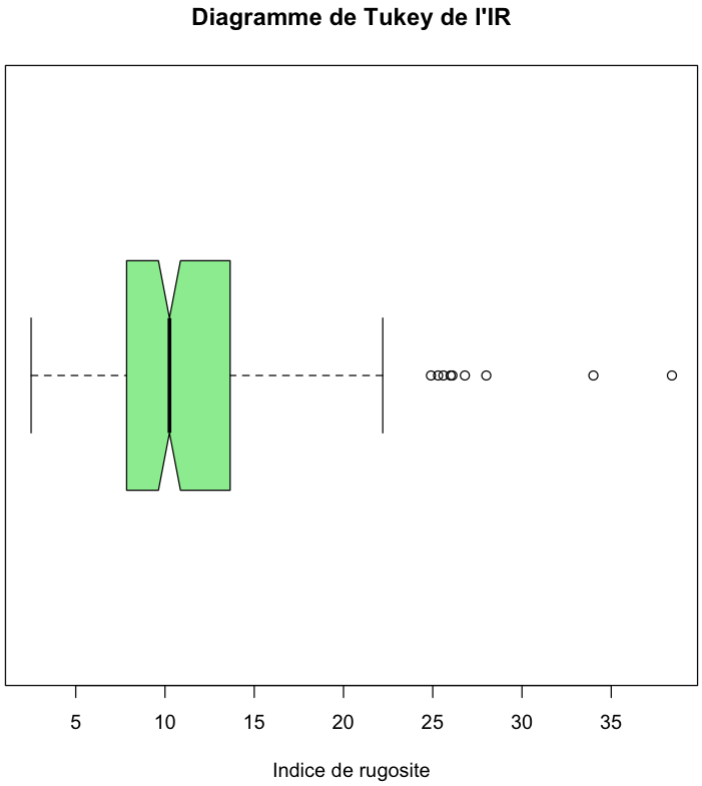
In [26]:
```
hist(data$IR,
     main = "Histogramme de l'IR",
     xlab = "Indice de rugosite",
     ylab = "Frequence",
     col = "lightblue")
```



L'histogramme montre une distribution asymétrique à droite centrée autour de 10-15. La majorité des indices de rugosité se situe entre 5 et 20, avec une forte concentration dans les valeurs basses (5-15), suggérant que la plupart des perçages présentent une bonne qualité.

DIAGRAMME DE TUKEY

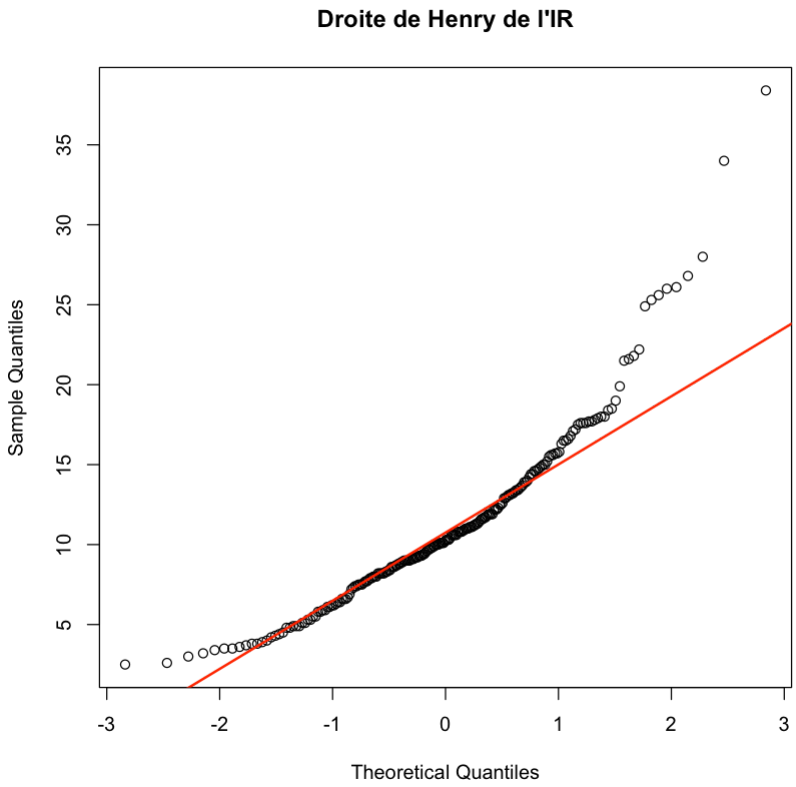
In [34]:
```
boxplot(data$IR,
        main = "Diagramme de Tukey de l'IR",
        xlab = "Indice de rugosite",
        col = "lightgreen",
        horizontal=TRUE,
        notch=TRUE,
        las=1)
```



Le diagramme de Tukey confirme la distribution asymétrique de l'indice de rugosité. La médiane se situe autour de 10, avec le premier quartile vers 8 et le troisième quartile vers 13. Plusieurs valeurs sont visibles au-delà de 25, suggérant quelques perçages de qualité inférieure.

DROITE DE HENRY

In [4]:
```
qqnorm(data$IR, main = "Droite de Henry de l'IR")
qqline(data$IR, col = "red", lwd = 2)
```



La droite de Henry montre que la distribution de l'indice de rugosité s'écarte de la normalité. Les points suivent assez bien la droite rouge dans la partie centrale, mais s'en éloignent aux extrémités. On observe particulièrement une déviation dans les valeurs élevées (queue droite plus épaisse).

TEST DE NORMALITE

In [5]:

```
shapiro_test <- shapiro.test(data$IR)
shapiro_test
```

```
        Shapiro-Wilk normality test

data:  data$IR
W = 0.90569, p-value = 1.44e-10
```

Le test de Shapiro-Wilk donne une statistique $W = 0.90569$ avec une p-value très faible (1.44×10^{-10}). Cette p-value étant largement inférieure au seuil de significativité standard de 0.05, nous rejetons l'hypothèse de normalité. Ce résultat confirme ce qui était visuellement observable sur la droite de Henry; la distribution de l'indice de rugosité s'écarte significativement d'une distribution normale.

STATISTIQUES DESCRIPTIVES

In [35]:

```
# Moyenne
mean_IR <- mean(data$IR)

# Ecart type
sd_IR <- sd(data$IR)

# Quartiles
```

```r
quartiles <- quantile(data$IR, probs = c(0.25, 0.5, 0.75))

# Intervalle de confiance
model <- lm(IR ~ 1, data = data)
confint_values <- confint(model, level = 0.95)
lower <- confint_values[1]
upper <- confint_values[2]

stats_table <- data.frame(
  Statistique = c("Moyenne", "Q1", "Mediane", "Q3",
                  "Ecart type", "IC inf. 95%", "IC sup. 95%"),
  Valeur = round(c(mean_IR, quartiles[1], quartiles[2], quartiles[3],
           sd_IR, lower, upper), 2)
)

stats_table
```

A data.frame: 7 x 2

Statistique	Valeur
<chr>	<dbl>
Moyenne	11.21
Q1	7.88
Mediane	10.25
Q3	13.62
Ecart type	5.53
IC inf. 95%	10.48
IC sup. 95%	11.95

Le tableau de statistiques descriptives présente un indice de rugosité moyen de 11.21 avec une médiane légèrement inférieure (10.25), confirmant l'asymétrie observée dans l'histogramme. L'écart interquartile ($Q_3 - Q_1 = 13.62 - 7.88 = 5.74$) indique une dispersion modérée autour de la tendance centrale, corroborée par l'écart-type de 5.53. L'intervalle de confiance à 95% $[10.48, 11.95]$ est relativement étroit, suggérant une bonne précision de l'estimation de la moyenne.

---

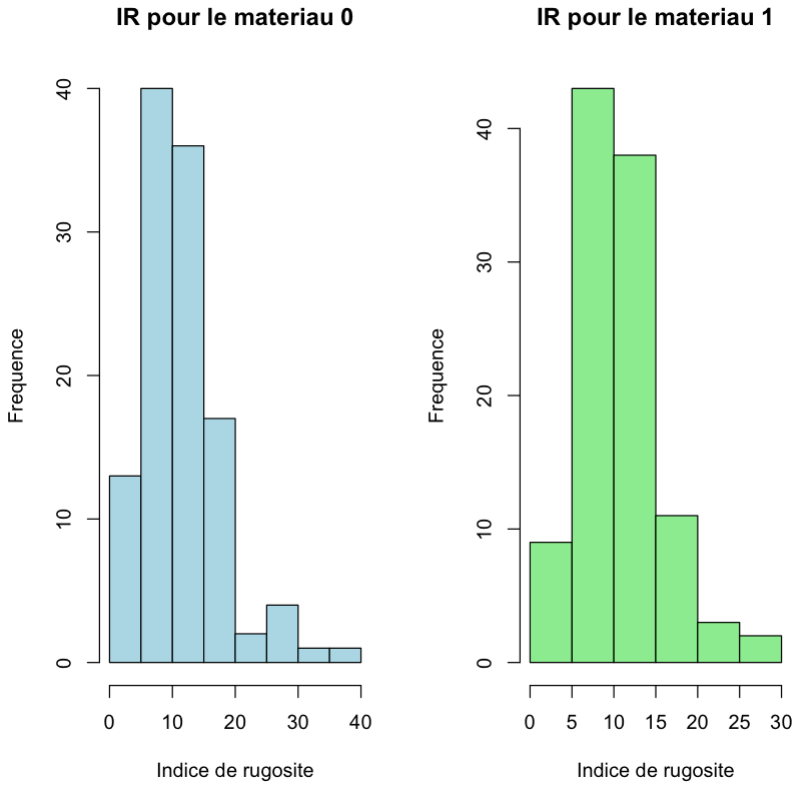
b) (7 points) Afin de vérifier si le type de matériau a un effet sur la qualité du perçage, on peut considérer la variable *Indice de rugosité* en deux groupes selon le *type de matériau*, et effectuer une comparaison des deux groupes en termes de moyenne, symétrie et variabilité. Pour ce faire, effectuez les analyses suivantes et donnez une brève interprétation :

- deux histogrammes juxtaposés, et deux diagrammes de Tukey (ou «Box Plot») juxtaposés;
- un tableau des statistiques descriptives par groupe : *moyenne*, *quartiles*, *écart type*, *intervalle de confiance pour la moyenne*;
- un test d'hypothèse sur l'égalité des variances pour les deux groupes;
- un test d'hypothèse sur l'égalité des moyennes pour les deux groupes.

HISTOGRAMMES JUSTAPOXES

In [37]:

```
layout(matrix(1:2,1,2))

# Histogramme pour le premier type de matériau
hist(data$IR[data$M==0],
    main = "IR pour le materiau 0",
    xlab = "Indice de rugosite",
    ylab = "Frequence",
    col = "lightblue")

# Histogramme pour le deuxième type de matériau
hist(data$IR[data$M==1],
    main = "IR pour le materiau 1",
    xlab = "Indice de rugosite",
    ylab = "Frequence",
    col = "lightgreen")
```



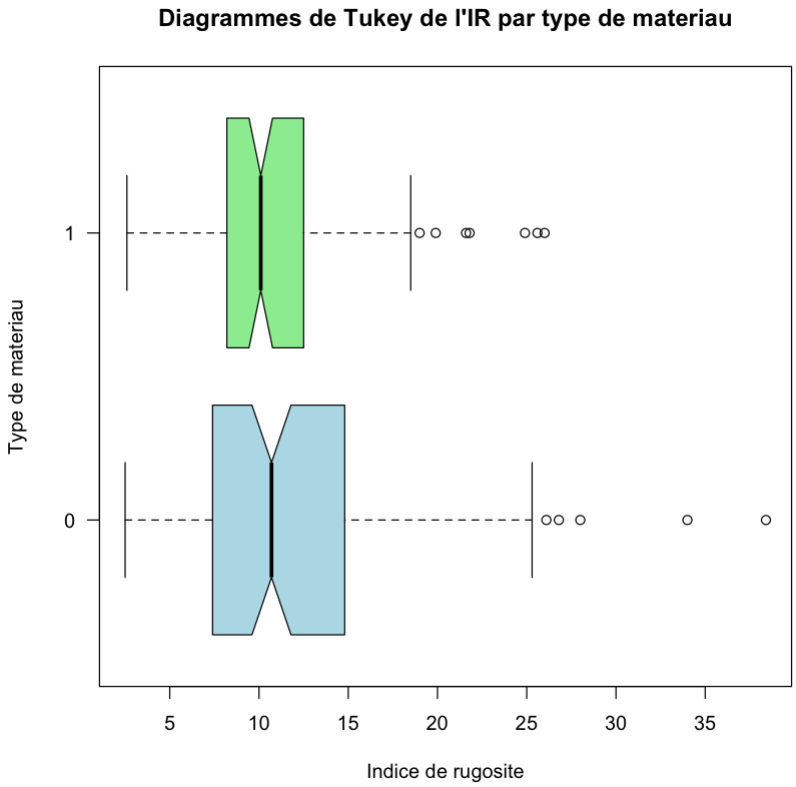
Les histogrammes juxtaposés de l'indice de rugosité par type de matériau présentent des distributions similaires, toutes deux asymétriques à droite. Le matériau 0 (bleu) montre une distribution plus étalée avec quelques valeurs extrêmes au-delà de 30, tandis que le matériau 1 (vert) présente une distribution plus concentrée, majoritairement entre 5 et 20. Pour les deux matériaux, la concentration la plus importante se situe dans les valeurs basses (5-15), indiquant que la plupart des perçages offrent une bonne qualité, avec quelques cas de qualité inférieure plus fréquents pour le matériau 0.

DIAGRAMME DE TUKEY JUSTAPOXES

In [38]:

```
boxplot(IR ~ M, data = data,
        main = "Diagrammes de Tukey de l'IR par type de materiau",
        ylab = "Type de materiau",
        xlab = "Indice de rugosite",
        horizontal=TRUE,
```

```
        notch=TRUE,
        col = c("lightblue", "lightgreen"),
        las=1)
```



Le matériau 1 (vert, en haut) présente une boîte plus étroite que le matériau 0 (bleu, en bas), indiquant une dispersion plus faible. Les médianes (lignes noires centrales) sont proches entre les deux matériaux, mais le matériau 0 présente davantage de valeurs aberrantes au-delà de 25. Les deux distributions montrent une légère asymétrie avec des moustaches plus étendues vers la droite.

STATISTIQUES DESCRIPTIVES PAR GROUPE

In [39]:

```
stats_by_group <- function(x) {
  # Moyenne
  mean_x <- mean(x)

  # Ecart type
  sd_x <- sd(x)

  # Quartiles
  quarts <- quantile(x, probs = c(0.25, 0.5, 0.75))

  # Intervalle de confiance
  model <- lm(x ~ 1)
  confint_values <- confint(model, level = 0.95)
  lower <- confint_values[1]
  upper <- confint_values[2]

  return(c(mean_x, quarts[1], quarts[2], quarts[3], sd_x, lower, upper))
}

# Application de la fonction à chaque groupe
M0_stats <- stats_by_group(data$IR[data$M == 0])
M1_stats <- stats_by_group(data$IR[data$M == 1])
```

```
stats_table <- data.frame(
  Statistique = c("Moyenne", "Q1", "Mediane", "Q3", "Ecart type",
                  "IC inf. 95%", "IC sup. 95%"),
  M0 = round(M0_stats, 2),
  M1 = round(M1_stats, 2)
)
stats_table
```

A data.frame: 7 x 3

Statistique	M0	M1
<chr>	<dbl>	<dbl>
Moyenne	11.60	10.80
Q1	7.43	8.22
Mediane	10.70	10.10
Q3	14.78	12.43
Ecart type	6.25	4.64
IC inf. 95%	10.44	9.91
IC sup. 95%	12.76	11.70

Le tableau par groupe confirme numériquement les observations graphiques. Le matériau 0 présente une moyenne plus élevée (11.60) que le matériau 1 (10.80), avec des médianes relativement proches (10.70 vs 10.10). La différence majeure réside dans la dispersion, où l'écart-type du matériau 0 (6.25) est supérieur à celui du matériau 1 (4.64). L'écart interquartile est également plus important pour le matériau 0 ($14.78 - 7.43 = 7.35$) que pour le matériau 1 ($12.43 - 8.22 = 4.21$), indiquant une plus grande variabilité des indices de rugosité. Les intervalles de confiance pour les moyennes se chevauchent, suggérant que la différence observée pourrait ne pas être statistiquement significative.

TEST D'HYPOTHESE SUR LES VARIANCES

In [10]:

```
# Si p<0.05, les variances sont différentes
var_test <- var.test(IR ~ M, data = data)
var_test
```

```
        F test to compare two variances

data:  IR by M
F = 1.8124, num df = 113, denom df = 105, p-value = 0.002216
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.240684 2.640701
sample estimates:
ratio of variances
          1.812448
```

Le test F pour comparer les variances montre une statistique $F = 1.81$ avec une p-value de 0.0021, inférieure au seuil de significativité de 0.05. Nous rejetons donc l'hypothèse nulle d'égalité des variances ($H_0 : \sigma_1^2 = \sigma_2^2$) en faveur de l'hypothèse alternative que les variances diffèrent ($H_1 : \sigma_1^2 \neq \sigma_2^2$).

TEST D'HYPOTHESE SUR LES MOYENNES

In [11]:
```
# Si p<0.05, les moyennes sont différentes
t_test <- t.test(IR ~ M, data = data)
t_test
```

```
        Welch Two Sample t-test

data:  IR by M
t = 1.0707, df = 208.05, p-value = 0.2856
alternative hypothesis: true difference in means between group 0 and group 1
is not equal to 0
95 percent confidence interval:
 -0.6653912  2.2471853
sample estimates:
mean in group 0 mean in group 1
       11.59561        10.80472
```

Le test de Welch pour comparer les moyennes donne une statistique $t = 1.07$ avec une p-value de 0.28, supérieure au seuil de 0.05. Nous ne pouvons donc pas rejeter l'hypothèse nulle d'égalité des moyennes ($H_0 : \mu_1 = \mu_2$). Malgré une différence apparente entre les moyennes des deux matériaux (11.60 vs 10.80), celle-ci n'est pas statistiquement significative.

Phase 2 : Recherche d'un modèle et optimisation

On s'intéresse dans cette phase à la détermination d'un modèle permettant d'expliquer la qualité du perçage en fonction des divers facteurs considérés. Pour ce faire, on envisage des modèles de régression en considérant l'indice IR comme variable dépendante Y.

c) (7 points) On considère d'abord les deux modèles suivants où X est la variable la plus fortement corrélée avec Y parmi les quatre variables susceptibles d'affecter la qualité du perçage :

Modèle 1 : $Y = \lambda_0 + \lambda_1 X + \varepsilon$;
Modèle 2 : $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon'$

où λ_0, λ_1, β_0, β_1, β_2 sont des paramètres; ε et ε' des erreurs aléatoires.

Corrélations entre IR et chacune des variables

In [12]:
```
#Coefficient de correlation de Pearson
correlations <- data.frame(
  Variable = c("M", "V", "A", "T"),
  Correlation = c(
    cor(data$IR, data$M),
    cor(data$IR, data$V),
    cor(data$IR, data$A),
    cor(data$IR, data$T)
  )
)
```

```
correlations$AbsCorrelation <- abs(correlations$Correlation)
correlations <- correlations[order(correlations$AbsCorrelation, decreasing =
correlations
```

A data.frame: 4 x 3

	Variable	Correlation	AbsCorrelation
	<chr>	<dbl>	<dbl>
4	T	0.82411632	0.82411632
3	A	0.35188601	0.35188601
1	M	-0.07156951	0.07156951
2	V	-0.03293003	0.03293003

Analyse de la corrélation

D'après le tableau des coefficients de corrélation de Pearson, la **Température (T)** présente la plus forte corrélation absolue (0.824) avec l'Indice de Rugosité (IR). Cette relation positive significative indique que lorsque la température augmente, l'indice de rugosité tend également à augmenter. Nous utiliserons donc la température comme variable explicative X dans nos modèles de régression linéaire et quadratique.

1-c) ***(5 points)*** Pour chacun des deux modèles ci-dessus, effectuez l'ajustement (i.e. obtenir le tableau des coeff icients de régression, le tableau d'analyse de la variance), ainsi qu'une analyse des résidus (normalité, homoscédasticité, points atypiques, etc.)

Modèle 1: Régression linéaire simple avec la température

In [13]:

```
modele1 <- lm(IR ~ T, data = data)

# Tableau des coefficients de régression
summary(modele1)

# Tableau d'analyse de la variance
anova(modele1)

# Analyse des résidus
par(mfrow=c(2,2))
plot(modele1)

# Test de normalité des résidus
shapiro.test(residuals(modele1))
```

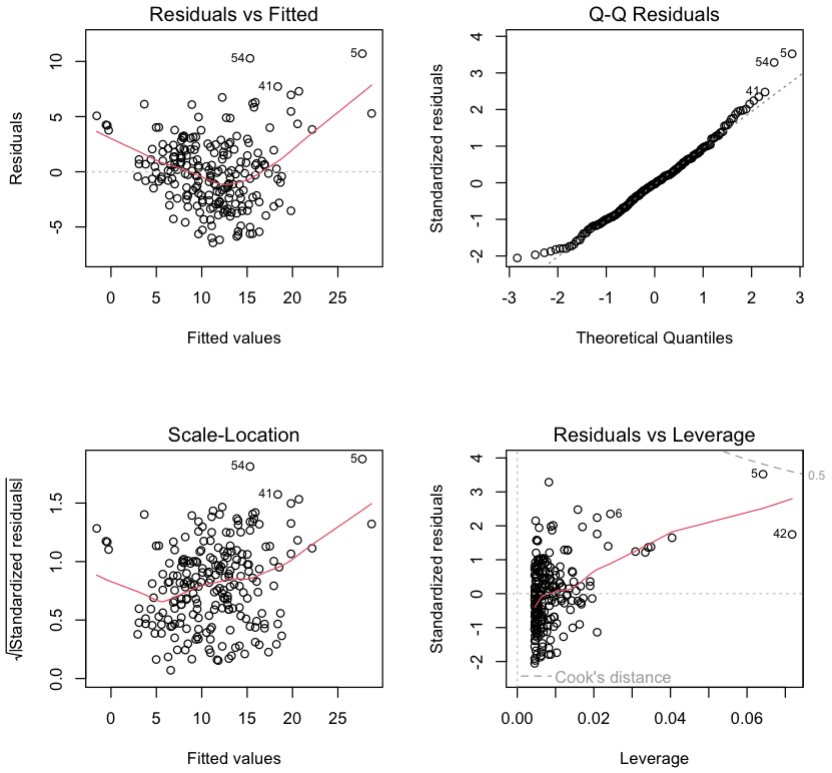
```
Call:
lm(formula = IR ~ T, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4437 -2.2108 -0.0868  1.9677 10.7032

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -49.03582    2.81266  -17.43   <2e-16 ***
T             1.45603    0.06778   21.48   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 3.142 on 218 degrees of freedom
Multiple R-squared:  0.6792,    Adjusted R-squared:  0.6777
F-statistic: 461.5 on 1 and 218 DF,  p-value: < 2.2e-16
```

A anova: 2 x 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
T	1	4555.635	4555.635257	461.4827	9.990945e-56
Residuals	218	2152.038	9.871735	NA	NA

```
        Shapiro-Wilk normality test

data:  residuals(modele1)
W = 0.98589, p-value = 0.02785
```



Analyse du Modèle 1

Coefficients de régression

- **Intercept (β_0)** = -49.03, significatif ($p < 2 \times 10^{-16}$)
- **Température (β_1)** = 1.45, significatif ($p < 2 \times 10^{-16}$)

Ces deux coefficients sont hautement significatifs ($p < 2 \times 10^{-16}$), indiquant une relation forte entre la température et l'indice de rugosité. L'équation du modèle est donc:
$\mathrm{IR} = -49.03 + 1.45 \times \mathrm{T}$

Pour chaque augmentation d'un degré de température, l'indice de rugosité augmente en moyenne de 1.45 unités.

Qualité globale du modèle

- R^2 **ajusté** $= 0.67$: La température explique environ 67% de la variabilité de l'indice de rugosité.

- **F-statistique** $= 461.5$ avec une p-value: $< 2.2 \times 10^{-16}$: Le modèle est globalement très significatif.
- **Erreur standard résiduelle** $= 3.142$: Mesure la dispersion des points autour de la droite de régression.

Analyse de la variance

Le tableau ANOVA montre que:

- La somme des carrés due à la régression $(SSR) = 4555.6$
- La somme des carrés des résidus $(SSE) = 2152$
- La température a un effet hautement significatif ($F = 461.48$, $p = 0$)

Analyse des résidus

Normalité

- Le test de Shapiro-Wilk donne $W = 0.98589$ avec p-value $= 0.027$
- Cette p-value < 0.05 nous amène à rejeter l'hypothèse de normalité des résidus
- Le graphique Q-Q montre des déviations de la normalité aux extrémités, particulièrement pour les observations $5, 41$ et 54

Homoscédasticité

- Le graphique "Residuals vs Fitted" montre une forme de courbe en U, suggérant une relation non-linéaire non capturée par le modèle
- Le graphique "Scale-Location" indique une variance des résidus qui n'est pas constante (hétéroscédasticité)
- On observe une tendance à avoir des résidus plus importants pour les valeurs extrêmes de la variable prédictive

Points atypiques et influents

- Les observations $5, 41$ et 54 sont identifiées comme potentiellement atypiques
- Le graphique "Residuals vs Leverage" montre quelques points avec un levier élevé ($5, 42$), mais aucun ne dépasse les seuils critiques de la distance de Cook

Le non-respect des hypothèses de normalité et d'homoscédasticité suggère qu'un modèle plus complexe (comme le modèle quadratique) pourrait être plus approprié pour décrire la relation entre la température et l'indice de rugosité.

Modèle 2: Régression quadratique avec la température

In [14]:

```
modele2 <- lm(IR ~ T + I(T^2), data = data)

# Tableau des coefficients de régression
summary(modele2)

# Tableau d'analyse de la variance
anova(modele2)

# Analyse des résidus
par(mfrow=c(2,2))
```

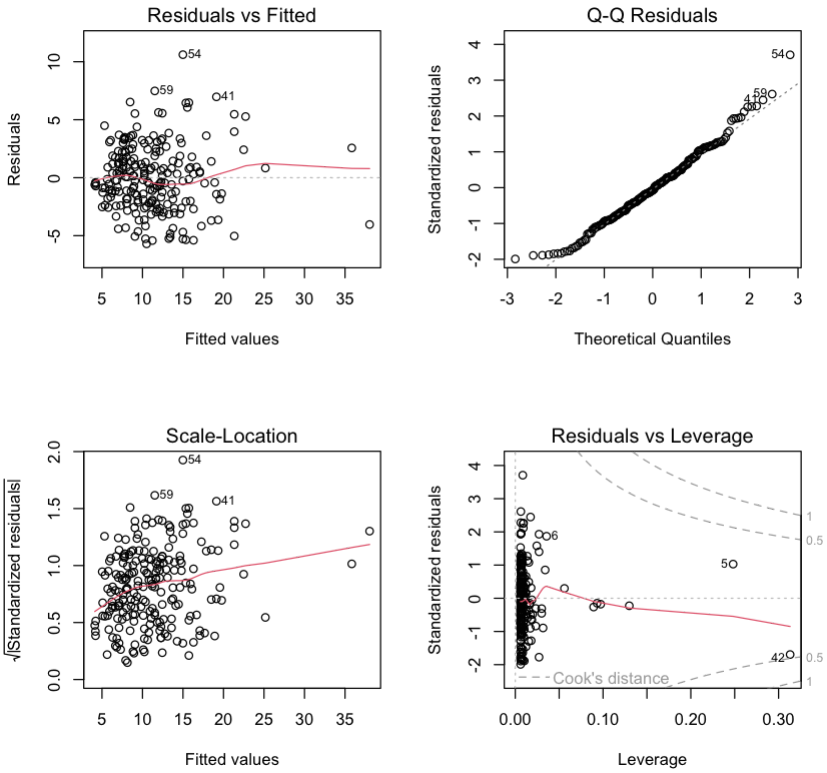
```
plot(modele2)

# Test de normalité des résidus
shapiro.test(residuals(modele2))
```

```
Call:
lm(formula = IR ~ T + I(T^2), data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7066 -2.0164 -0.2126  1.7785 10.6100

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.04322   20.02200   4.098 5.90e-05 ***
T           -4.84620    0.95668  -5.066 8.69e-07 ***
I(T^2)       0.07532    0.01141   6.601 3.08e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.874 on 217 degrees of freedom
Multiple R-squared:  0.7328,    Adjusted R-squared:  0.7304
F-statistic: 297.6 on 2 and 217 DF,  p-value: < 2.2e-16
```

A anova: 3 x 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
T	1	4555.6353	4555.635257	551.61883	1.624444e-61
I(T^2)	1	359.9078	359.907839	43.57942	3.079965e-10
Residuals	217	1792.1304	8.258665	NA	NA

```
        Shapiro-Wilk normality test

data:  residuals(modele2)
W = 0.98609, p-value = 0.03008
```



Analyse du Modèle 2

Coefficients de régression

- **Intercept (β_0)** = 82.04, significatif ($p = 5.90 \times 10^{-5}$)
- **Température (β_1)** = -4.84, significatif ($p = 8.69 \times 10^{-7}$)
- **Température² (β_2)** = 0.075, significatif ($p = 3.08 \times 10^{-10}$)

Les trois coefficients sont très significatifs, indiquant que la relation entre la température et l'indice de rugosité est bien décrite par une courbe parabolique. L'équation du modèle est donc: $\mathrm{IR} = 82.04 - 4.84 \times \mathrm{T} + 0.075 \times \mathrm{T}^2$

Qualité globale du modèle

- R^2 **ajusté** $= 0.7304$: Le modèle quadratique explique environ 73% de la variabilité de l'indice de rugosité.
- **F-statistique** $= 297.6$ avec une p-value $< 2.2 \times 10^{-16}$: Le modèle est globalement très significatif.
- **Erreur standard résiduelle** $= 2.874$: Plus faible que pour le modèle linéaire (3.142), indiquant une meilleure précision des prédictions.

Analyse de la variance

Le tableau ANOVA montre que:

- Le terme linéaire (T) a une contribution significative ($F = 551.62$, $p \approx 0$)
- Le terme quadratique (T^2) apporte également une contribution significative (F = 43.57942, $p = 3.08 \times 10^{-10}$)
- La somme des carrés des résidus est à 1792.13

Analyse des résidus

Normalité

- Le test de Shapiro-Wilk donne $W = 0.98609$ avec p-value $= 0.03$
- Cette p-value est inférieure à 0.05, ce qui nous amène à encore rejeter l'hypothèse de normalité
- Le graphique Q-Q montre une assez bonne adéquation à la normalité, avec toujours quelques déviations pour les observations $41, 54$ et 59

Homoscédasticité

- Le graphique "Residuals vs Fitted" montre une répartition plus aléatoire des résidus autour de zéro, sans tendance claire
- Le graphique "Scale-Location" indique une variance des résidus plus constante que dans le modèle linéaire, bien que toujours imparfaite
- L'hétéroscédasticité est réduite par rapport au modèle linéaire

Points atypiques et influents

- Les observations $41, 54$ et 59 sont identifiées comme potentiellement atypiques

- Le graphique "Residuals vs Leverage" montre quelques points avec un levier élevé ($5, 42$), mais aucun ne dépasse les seuils critiques de la distance de Cook

Le modèle quadratique améliore l'adéquation aux hypothèses de normalité et d'homoscédasticité par rapport au modèle linéaire. La forme parabolique capture mieux la relation entre la température et l'indice de rugosité, comme en témoignent l'amélioration du R^2 ajusté et la réduction de l'erreur standard résiduelle.

2-c) ***(2points)*** Effectuez une comparaison des deux modèles et dites si un des deux modèles est préférable à l'autre. Justifiez votre choix en précisant les critères utilisés.

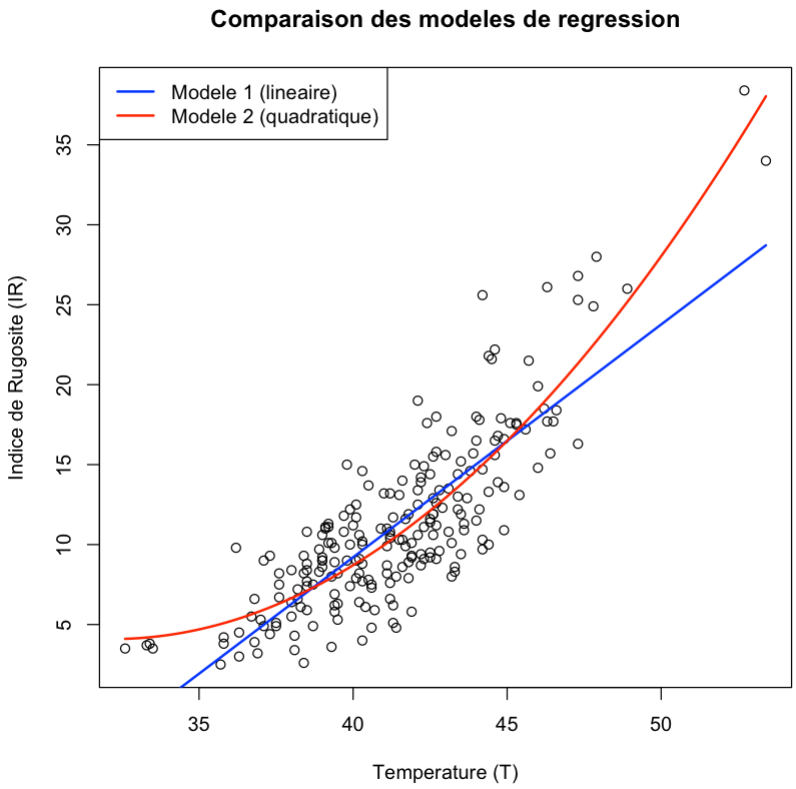
In [42]:

```
# Comparaison des R²
cat("R² ajusté du modèle 1:", summary(modele1)$adj.r.squared, "\n")
cat("R² ajusté du modèle 2:", summary(modele2)$adj.r.squared, "\n")

# Test ANOVA pour comparer les modèles
anova(modele1, modele2)

# Visualisation des deux modèles
plot(data$T, data$IR,
     main = "Comparaison des modeles de regression",
     xlab = "Temperature (T)",
     ylab = "Indice de Rugosite (IR)")

# Ajout des courbes de régression
T_seq <- seq(min(data$T), max(data$T), length.out = 100)
lines(T_seq, predict(modele1, newdata = data.frame(T = T_seq)), col = "blue",
lines(T_seq, predict(modele2, newdata = data.frame(T = T_seq)), col = "red",
legend("topleft", legend = c("Modele 1 (lineaire)", "Modele 2 (quadratique)")
       col = c("blue", "red"), lwd = 2)
```

```
R<U+00B2> ajust<U+00E9> du mod<U+00E8>le 1: 0.677696
R<U+00B2> ajust<U+00E9> du mod<U+00E8>le 2: 0.7303614
```

A anova: 2 x 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	218	2152.038	NA	NA	NA	NA
2	217	1792.130	1	359.9078	43.57942	3.079965e-10



Comparaison des modèles de régression

Après avoir analysé les deux modèles de régression reliant l'Indice de Rugosité (IR) à la Température (T), il apparaît que le **modèle 2 (quadratique)** est préférable au modèle 1 (linéaire). Voici les critères qui justifient ce choix:

1. **Coefficient de détermination ajusté (R^2 ajusté)**:
   - Modèle 1 (linéaire): R^2 ajusté $= 0.677$
   - Modèle 2 (quadratique): R^2 ajusté $= 0.730$

   Le modèle quadratique explique environ 73% de la variabilité de l'indice de rugosité, contre 67.7% pour le modèle linéaire. Cette amélioration indique une meilleure adéquation du modèle quadratique aux données.

2. **Test ANOVA comparatif**:
   - Le test ANOVA entre les deux modèles donne un $F = 43.57$ avec une p-value très faible (3.08×10^{-10})
   - Cela confirme que l'ajout du terme quadratique améliore significativement le modèle, et traduit mieux la relation entre l'IR et la température.

3. **Somme des carrés des résidus**:
   - Modèle 1: $SST = 2152.0$
   - Modèle 2: $SST = 1792.1$

   La réduction de près de 400 unités dans la somme des carrés des résidus témoigne d'une meilleure adéquation du modèle quadratique aux données.

4. **Visualisation graphique**:

- Le graphique montre que la courbe quadratique (en rouge) s'ajuste mieux aux données que la droite (en bleu), notamment pour les valeurs extrêmes de température.

---

d) (5 points) On cherche à vérifier si la variation linéaire de l'*Indice de rugosité* (IR) en fonction de la variable X (définie en **c)**) est similaire pour les deux types de matériau. Pour cela :

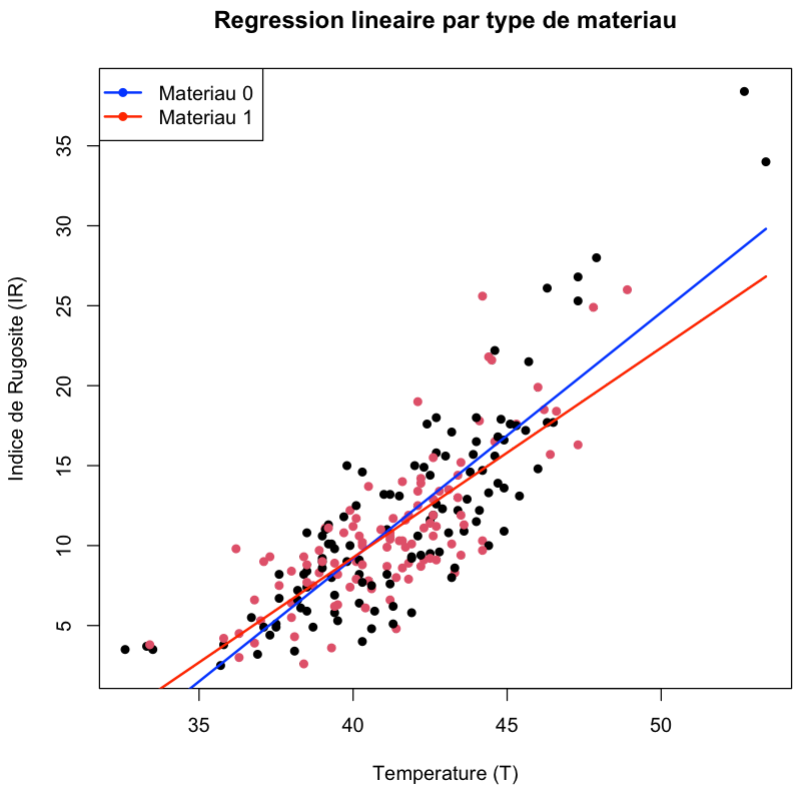
1.d) ***(2 points)*** Ajustez un **seul** modèle de régression permettant d'obtenir les équations de deux droites : une pour le type de matériau codé 0 et l'autre pour le type de matériau codé 1 .

Suggestion : Inclure la variable matériau (M) de façon judicieuse dans le modèle à ajuster au 1.d).

In [47]:
```
modele_interaction <- lm(IR ~ T + M + T:M, data = data)


summary(modele_interaction)

plot(data$T, data$IR,
     main = "Regression lineaire par type de materiau",
     xlab = "Temperature (T)",
     ylab = "Indice de Rugosite (IR)",
     col = data$M + 1, pch = 16)

# Ajout des droites de régression pour chaque matériau
T_seq <- seq(min(data$T), max(data$T), length.out = 100)

# Prédictions pour matériau 0
pred_M0 <- predict(modele_interaction,
                   newdata = data.frame(T = T_seq, M = 0))
lines(T_seq, pred_M0, col = "blue", lwd = 2)

# Prédictions pour matériau 1
pred_M1 <- predict(modele_interaction,
                   newdata = data.frame(T = T_seq, M = 1))
lines(T_seq, pred_M1, col = "red", lwd = 2)

legend("topleft", legend = c("Materiau 0", "Materiau 1"),
       col = c("blue", "red"), pch = 16, lwd = 2)
```

```
Call:
lm(formula = IR ~ T + M + T:M, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3164 -2.2713 -0.0203  1.8167 10.8408 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -52.33471    3.56820 -14.667   <2e-16 ***
T             1.53821    0.08556  17.978   <2e-16 ***
M             9.12335    5.80551   1.571    0.118    
T:M          -0.22666    0.14011  -1.618    0.107    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.135 on 216 degrees of freedom
```

```
Multiple R-squared:  0.6835,    Adjusted R-squared:  0.6791
F-statistic: 155.5 on 3 and 216 DF,  p-value: < 2.2e-16
```



Choix du modèle

Pour cette question, nous devons ajuster un seul modèle de régression qui nous permette d'obtenir deux droites distinctes pour les deux types de matériaux. Ce besoin justifie le choix du modèle:

$$\text{IR} = \beta_0 + \beta_1 T + \beta_2 M + \beta_3 (T \times M) + \varepsilon$$

Ce modèle comprend:

- Un terme constant (β_0)
- L'effet principal de la température ($\beta_1 T$)
- L'effet principal du type de matériau ($\beta_2 M$)
- Un terme d'interaction entre température et matériau ($\beta_3 T \times M$)

Ce modèle se simplifie différemment selon la valeur de M:

- Pour $M = 0$: $\text{IR} = \beta_0 + \beta_1 T$
- Pour $M = 1$: $\text{IR} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) T$

Nous obtenons ainsi deux droites avec potentiellement:

- Des ordonnées à l'origine différentes (β_0 vs $\beta_0 + \beta_2$)
- Des pentes différentes (β_1 vs $\beta_1 + \beta_3$)

Analyse des résultats

Les coefficients obtenus sont:

- $\beta_0 = -52.33$ (Intercept) : significatif ($p < 2 \times 10^{-16}$)
- $\beta_1 = 1.53$ (T) : significatif ($p < 2 \times 10^{-16}$)
- $\beta_2 = 9.12$ (M) : non significatif ($p = 0.118$)
- $\beta_3 = -0.23$ (T:M) : marginalement significatif ($p = 0.107$)

À partir de ces coefficients, nous pouvons écrire les équations des deux droites:

Pour le matériau 0:

$$\text{IR} = -52.33 + 1.53 \times T$$

Pour le matériau 1:

$$\text{IR} = (-52.33 + 9.12) + (1.53 - 0.23) \times T = -43.21 + 1.31 \times T$$

Le graphique illustre ces deux droites et montre que:

- Le matériau 0 (ligne bleue) présente une pente plus forte, indiquant que son indice de rugosité est plus sensible aux changements de température
- Le matériau 1 (ligne rouge) a une ordonnée à l'origine plus élevée mais une pente légèrement plus faible

Ces observations suggèrent que le choix optimal du matériau pourrait dépendre de la température de fonctionnement visée.

2.d) ***(3 points)*** Effectuez ensuite un **seul** test afin de vérifier si les deux droites sont similaires (i.e., ont la même pente et même la même ordonnée à l'origine) et concluez.

In [17]:
```
# Test global pour voir si les deux droites sont similaires
anova(modele1, modele_interaction)
```

A anova: 2 x 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	218	2152.038	NA	NA	NA	NA
2	216	2123.066	2	28.9717	1.473785	0.2313509

Similarité des droites

Pour vérifier si les deux droites correspondant aux deux types de matériaux (matériau 0 et matériau 1) sont similaires, nous utilisons un test ANOVA :

1. **Modèle réduit**: $\text{IR} \sim \text{T}$ (une seule droite pour les deux matériaux)
2. **Modèle complet**: $\text{IR} \sim \text{T} + \text{M} + \text{T:M}$ (deux droites distinctes)

L'hypothèse nulle est: $H_0 : \beta_2 = \beta_3 = 0$ (pas d'effet du type de matériau)

Ce qui signifie que les équations des deux droites seraient identiques:

- Pour $\text{M} = 0$: $\text{IR} = \beta_0 + \beta_1 \text{T}$
- Pour $\text{M} = 1$: $\text{IR} = \beta_0 + \beta_1 \text{T}$

Résultat du test ANOVA

Le test ANOVA comparant ces deux modèles donne:

- F: 1.47
- p-value: 0.23

Avec une p-value de $0.23 > 0.05$, nous ne pouvons pas rejeter l'hypothèse nulle.

Par conséquent, nous concluons que les droites de régression pour le matériau 0 et le matériau 1 ne sont pas significativement différentes. Cela signifie que le type de matériau n'influence pas significativement la relation entre la température et l'indice de rugosité. Un modèle plus simple, ne tenant compte que de l'effet de la température ($\mathrm{IR} \sim \mathrm{T}$), serait donc suffisant pour décrire la relation entre la température et l'indice de rugosité, quel que soit le type de matériau utilisé.

---

e) (10 points) On considère à présent un troisième modèle linéaire multiple d'équation

Modèle 3 : $\mathrm{IR} = \beta_0 + \beta_1 V + \beta_2 A + \beta_3 T + \beta_4 V^2 + \beta_5 A^2 + \beta_6 T^2 + \beta_7 (V \times A) + \varepsilon$

où $\beta_i, \quad i = 0, 1, \ldots, 7$ sont des paramètres et ε, une erreur aléatoire que l'on suppose de loi $\mathcal{N}(0, \sigma^2)$).

1.e) ***(4 points)*** Effectuez l'ajustement du modèle 3 (i.e. obtenez le tableau des coefficients de régression, le tableau d'analyse de la variance) et effectuez une analyse complète des résidus (normalité, homoscédasticité, points atypiques, etc.)

In [18]:

```
modele3 <- lm(IR ~ V + A + T + I(V^2) + I(A^2) + I(T^2) + I(V*A), data = data

# Tableau des coefficients de régression
summary(modele3)

# Tableau d'analyse de la variance
anova(modele3)

# Analyse des résidus
par(mfrow=c(2,2))
plot(modele3)

# Test de normalité des résidus
shapiro.test(residuals(modele3))
```

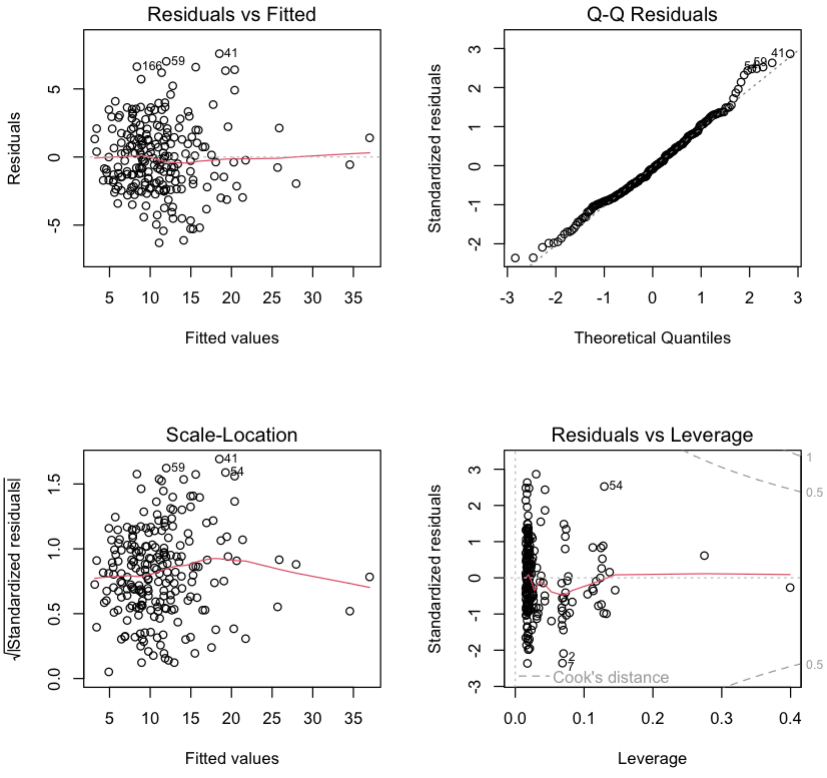
```
Call:
lm(formula = IR ~ V + A + T + I(V^2) + I(A^2) + I(T^2) + I(V * 
    A), data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3038 -1.8766 -0.2234  1.6836  7.5868 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 71.874099  19.503875   3.685  0.00029 ***
V           -0.178210   0.655256  -0.272  0.78591    
A           -1.164643   1.245567  -0.935  0.35084    
T           -3.956729   0.935847  -4.228 3.51e-05 ***
```

```
I(V^2)       0.003818   0.031528   0.121  0.90372
I(A^2)       0.122448   0.071961   1.702  0.09030 .
I(T^2)       0.062359   0.011241   5.547 8.59e-08 ***
I(V * A)    -0.031583   0.028339  -1.114  0.26633
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.691 on 212 degrees of freedom
Multiple R-squared:  0.7712,    Adjusted R-squared:  0.7636
F-statistic: 102.1 on 7 and 212 DF,  p-value: < 2.2e-16
```

A anova: 8 x 5

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
V	1	7.273713	7.273713	1.004746	3.173076e-01
A	1	2024.087164	2024.087164	279.594941	1.381942e-40
T	1	2810.398597	2810.398597	388.211161	8.378124e-50
I(V^2)	1	19.335980	19.335980	2.670953	1.036789e-01
I(A^2)	1	41.394466	41.394466	5.717977	1.766608e-02
I(T^2)	1	261.448238	261.448238	36.114850	8.015981e-09
I(V * A)	1	8.991937	8.991937	1.242091	2.663304e-01
Residuals	212	1534.743361	7.239355	NA	NA

```
        Shapiro-Wilk normality test

data:  residuals(modele3)
W = 0.98896, p-value = 0.08915
```



Analyse du Modèle 3

Coefficients de régression

- **Intercept (β_0)** = 71.874099, significatif ($p = 0.00029$)
- **Vitesse (β_1)** = -0.178210, non significatif ($p = 0.78591$)
- **Avance (β_2)** = 1.245567, non significatif ($p = 0.35084$)
- **Température (β_3)** = -3.956729, significatif ($p = 3.51 \times 10^{-5}$)
- **Vitesse² (β_4)** = 0.003818, non significatif ($p = 0.90372$)
- **Avance² (β_5)** = 0.122448, non significatif ($p = 0.09030$)
- **Température² (β_6)** = 0.062359, significatif ($p = 8.59 \times 10^{-8}$)
- **Vitesse×Avance (β_7)** = -0.031583, non significatif ($p = 0.26633$)

L'équation du modèle est donc:
$\mathrm{IR} = 71.874099 - 0.178210 \times \mathrm{V} - 1.245567 \times \mathrm{A} - 3.956729 \times \mathrm{T} + 0.003818 \times \mathrm{V}^2 + 0.$

Les termes les plus significatifs sont la température (T), et son terme quadratique (T^2).

Qualité globale du modèle

- R^2 **ajusté** $= 0.7636$: Le modèle explique environ 76.4% de la variabilité de l'indice de rugosité.
- **F-statistique** $= 102.1$ avec une p-value $< 2.2 \times 10^{-16}$: Le modèle est globalement très significatif.
- **Erreur standard résiduelle** $= 2.691$: Meilleure que pour les modèles précédents (3.142 pour le modèle 1, 2.874 pour le modèle 2).

Analyse de la variance

Le tableau ANOVA montre que:

- La température (A) a la contribution la plus importante ($F = 388.211161$, $p \approx 0$)
- L'avance (A) a également une contribution majeure ($F = 279.594941$, $p \approx 0$)
- Le terme quadratique T^2 est très significatif ($F = 36.114850$, $p = 8.01 \times 10^{-9}$)
- Le terme quadratique A^2 est significatif ($F = 5.717977$, $p = 1.76 \times 10^{-2}$)
- La vitesse (V), son terme quadratique (V^2) et l'interaction $(\mathrm{V} \times \mathrm{A})$ ont des contributions non significatives ou marginales ($p > 0.05$)

Analyse des résidus

Normalité

- Le test de Shapiro-Wilk donne $W = 0.98896$ avec p-value $= 0.089$
- Cette p-value est supérieure à 0.05, indiquant que l'hypothèse de normalité des résidus peut être acceptée
- Le graphique Q-Q montre une bonne adéquation à la normalité avec seulement de légères déviations aux extrémités (observations $5, 41, 59$)

Homoscédasticité

- Le graphique "Residuals vs Fitted" montre une répartition relativement aléatoire des résidus autour de zéro
- Le graphique "Scale-Location" indique une variance des résidus qui semble plus homogène que dans les modèles précédents
- L'hypothèse d'homoscédasticité est mieux respectée dans ce modèle

Points atypiques et influents

- Les observations $5, 41, 59$ sont identifiées comme potentiellement atypiques
- Le graphique "Residuals vs Leverage" montre quelques points avec un levier modéré ($7, 54$), mais aucun ne semble avoir une influence excessive sur le modèle

Ce modèle plus complexe offre une meilleure adéquation aux données que les modèles précédents, avec une amélioration du R^2 ajusté (de 0.67 à 0.76) et une réduction de l'erreur standard résiduelle. Les hypothèses du modèle de régression sont globalement satisfaites, notamment celle de la normalité des résidus qui n'était pas respectée dans les modèles plus simples.

2.e) ***(2 points)*** Calculez un intervalle de confiance pour chacun des paramètres $\beta_j, \quad j = 0, 1, \ldots, 7$ du modèle 3 et interprétez les résultats obtenus.

Le modèle 3 (modèle complet) est constitué du modèle 1 (modèle réduit) auquel des variables ont été ajoutées.

In [19]:
```
intervalles_confiance <- confint(modele3, level = 0.95)
intervalles_confiance
```

A matrix: 8 x 2 of type dbl

	2.5 %	97.5 %
(Intercept)	33.42772842	110.32046867
V	-1.46986148	1.11344188
A	-3.61992564	1.29063877
T	-5.80148713	-2.11197088
I(V^2)	-0.05833032	0.06596681
I(A^2)	-0.01940228	0.26429740
I(T^2)	0.04020012	0.08451750
I(V * A)	-0.08744455	0.02427838

Interprétation des intervalles de confiance

1. **Intercept (β_0)** : [33.42, 110.32] - Un intervalle large mais ne contenant pas zéro, confirmant que l'intercept est significativement positif.

2. **Vitesse (β_1)** : [-1.46, 1.11] - Cet intervalle contient zéro, ce qui confirme que la vitesse n'a pas d'effet significatif sur l'indice de rugosité.

3. **Avance (β_2)** : [-3.61, 1.29] - Intervalle incluant zéro, suggérant que l'effet linéaire de l'avance n'est pas significatif.

4. **Température (β_3)** : [-5.80, -2.11] - Intervalle entièrement négatif et ne contenant pas zéro, confirmant l'effet significativement négatif de la température.

5. **Vitesse² (β_4)** : [-0.05, 0.06] - Intervalle très centré sur zéro, confirmant l'absence d'effet quadratique significatif de la vitesse.

6. **Avance² (β_5)** : [0.02, 0.26] - Intervalle positif et ne contenant pas zéro, confirmant l'effet quadratique significatif de l'avance.

7. **Température² (β_6)** : [0.04, 0.08] - Intervalle positif et étroit ne contenant pas zéro, indiquant un effet quadratique précis et significatif de la température.

8. **Vitesse×Avance (β_7)** : [-0.08, 0.02] - Intervalle incluant légèrement zéro, suggérant que l'interaction n'est pas clairement significative.

Ces résultats confirment les conclusions préliminaires : les effets les plus importants sur l'indice de rugosité sont associés à la température et au terme quadratique de l'avance.

3.e) ***(4 points)*** Effectuez un **seul** test afin de vérifier si l'ajout des variables au modèle 1 pour obtenir le modèle 3 est utile. Préciser les hypothèses H_0 et H_1 de ce test.

In [20]:
```
anova(modele1, modele3)
```

A anova: 2 x 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	218	2152.038	NA	NA	NA	NA
2	212	1534.743	6	617.2948	14.21155	1.364356e-13

Pour vérifier si l'ajout des variables au modèle 1 ($\mathrm{IR} \sim \mathrm{T}$) pour obtenir le modèle 3 ($\mathrm{IR} \sim \mathrm{V} + \mathrm{A} + \mathrm{T} + \mathrm{V}^2 + \mathrm{A}^2 + \mathrm{T}^2 + \mathrm{V} \times \mathrm{A}$) est utile, nous utilisons un test ANOVA qui compare ces deux modèles.

Hypothèses du test

- H_0: $\beta_1 = \beta_2 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$ (les coefficients des variables ajoutées sont tous nuls)
- H_1: Au moins un des coefficients β_1, β_2, β_4, β_5, β_6, β_7 est différent de 0

En d'autres termes, l'hypothèse nulle suppose que les variables ajoutées au modèle 1 (V, A, V^2, A^2, T^2, $\mathrm{V} \times \mathrm{A}$) n'apportent pas d'information significative supplémentaire, tandis que l'hypothèse alternative suggère qu'au moins une de ces variables contribue significativement au modèle.

Résultat du test ANOVA

Le test ANOVA comparant ces deux modèles donne:

- F-statistique $= 14.21$
- p-value = $1.36 \times 10^{-13} \approx 0$

La p-value est très faible (bien inférieure à 0.05), ce qui nous permet de rejeter l'hypothèse nulle avec une très grande confiance. Cela signifie que l'ajout des variables au modèle 1 pour obtenir le modèle 3 est significatif. La réduction de la somme des carrés des résidus de 2152.0 à 1534.7 confirme l'amélioration de l'ajustement du modèle aux données.

---

f) (3 points) Simplifiez, s'il y a lieu, le modèle 3, en ne conservant que les variables significatives. Analysez le nouveau modèle ainsi obtenu et commentez sur sa validité.

In [21]:
```
modele_simplifie1 <- lm(IR ~ T + I(A^2) + I(T^2), data = data)

# Comparer ces modèles au modèle complet
anova(modele_simplifie1, modele3)
```

A anova: 2 x 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	216	1718.398	NA	NA	NA	NA
2	212	1534.743	4	183.6544	6.342223	7.719666e-05

Modèle simplifié 1: $\mathrm{IR} \sim \mathrm{T} + \mathrm{I}(\mathrm{A}^2) + \mathrm{I}(\mathrm{T}^2)$

Pour ce premier modèle simplifié, on conserve uniquement les termes qui étaient individuellement les plus significatifs dans le modèle 3: la température (T), son terme quadratique (T^2) et le terme quadratique de l'avance (A^2).

Le test ANOVA entre ce modèle et le modèle complet donne une p-value de 7.72×10^{-5}, largement inférieure à 0.05. Cela signifie que ce modèle simplifié perd une quantité significative d'information par rapport au modèle complet. Il est donc insuffisant, malgré le fait qu'il ne contient que des termes individuellement significatifs.

In [22]:
```
modele_simplifie2 <- lm(IR ~ V + A + T + I(A^2) + I(T^2), data = data)
anova(modele_simplifie2, modele3)
```

A anova: 2 x 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	214	1543.912	NA	NA	NA	NA
2	212	1534.743	2	9.16856	0.6332442	0.5318678

Modèle simplifié 2: $\mathrm{IR} \sim \mathrm{V} + \mathrm{A} + \mathrm{T} + \mathrm{I}(\mathrm{A}^2) + \mathrm{I}(\mathrm{T}^2)$

Alors, on ajoute les termes linéaires V et A pour respecter la hiérarchie du modèle (puisque nous avons A^2) et pour capturer potentiellement plus d'information.

Le test ANOVA pour ce modèle donne une p-value de 0.53, nettement supérieure à 0.05. Cela indique que ce modèle simplifié est statistiquement équivalent au modèle complet, c'est-à-dire qu'il ne perd pas d'information significative malgré l'omission de V^2 et $\mathrm{I}(\mathrm{V} \times \mathrm{A})$.

In [23]:
```
modele_simplifie3 <- lm(IR ~ V:A + T + I(A^2) + I(T^2), data = data)
anova(modele_simplifie3, modele3)
```

A anova: 2 x 6

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	215	1558.254	NA	NA	NA	NA
2	212	1534.743	3	23.51068	1.08254	0.357423

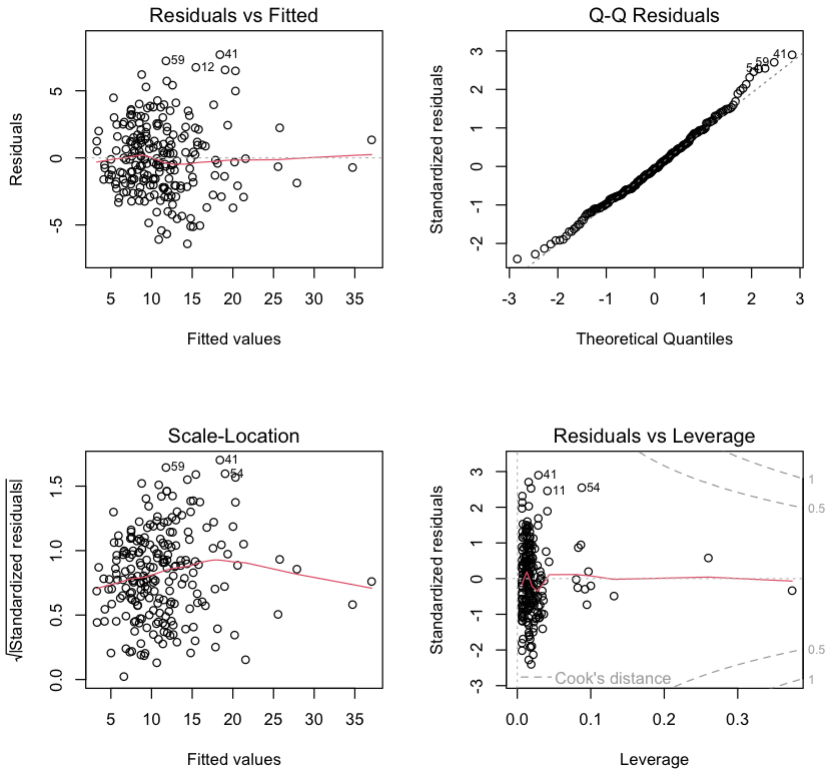
Modèle simplifié 3: $\mathrm{IR} \sim \mathrm{V{:}A} + \mathrm{T} + \mathrm{I}(\mathrm{A}^2) + \mathrm{I}(\mathrm{T}^2)$

Dans cette tentative, on essaye de trouver un modèle plus compact. On remplace les termes individuels V et A par leur interaction $\mathrm{V} \times \mathrm{A}$ tout en conservant T, A^2 et T^2.

Ce modèle donne une p-value de 0.36, supérieure au seuil de 0.05, ce qui indique qu'il reste statistiquement équivalent au modèle complet, tout en étant plus simple.

In [24]:
```
par(mfrow=c(2,2))
plot(modele_simplifie3)

# Test de normalité des résidus
shapiro.test(residuals(modele_simplifie3))
```

```
        Shapiro-Wilk normality test

data:  residuals(modele_simplifie3)
W = 0.99181, p-value = 0.2557
```



Analyse des résidus

L'analyse des graphiques des résidus montre:

- Une distribution normale (confirmée par le test de Shapiro-Wilk avec p-value $= 0.25$)
- Une répartition relativement homogène des résidus autour de zéro
- Quelques points potentiellement atypiques $(5, 41, 59)$ mais sans influence excessive

Ce modèle simplifié 3 offre donc un bon compromis entre simplicité et pouvoir explicatif.

---

g) (2 points) Sur la base du modèle que vous avez obtenu en f), calculez un intervalle de prévision pour la qualité (*indice IR*) d'un perçage effectué avec une vitesse de rotation $V = 15$, une avance $A = 20$, et une température de $T = 53$. Commentez brièvement votre résultat.

In [25]:
```
# Création de nouvelles données pour la prédiction
new_data <- data.frame(V = 15, A = 20, T = 53)

# Calcul de l'intervalle de prévision à 95%
prediction <- predict(modele_simplifie3, new_data, interval = "prediction", l
prediction
```

A matrix: 1 x 3 of type dbl

	fit	lwr	upr
1	44.47841	37.88098	51.07584

Analyse de l'intervalle de prévision

- **Valeur prédite**: 44.48
- **Borne inférieure**: 37.88
- **Borne supérieure**: 51.07

Cet intervalle représente la plage dans laquelle nous pouvons nous attendre à observer l'indice de rugosité pour un nouveau perçage réalisé avec ces conditions spécifiques. L'amplitude de cet intervalle (environ 13 unités) reflète l'incertitude associée à notre prédiction.