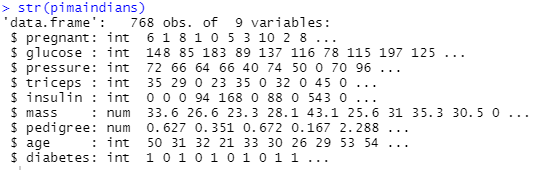
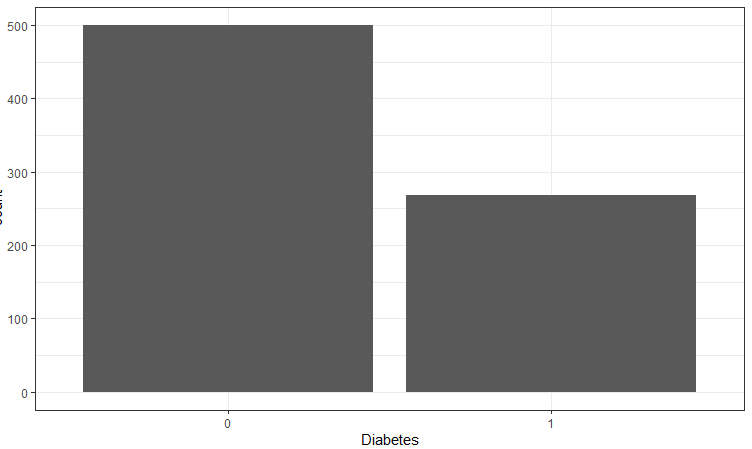
**Compte rendu TP noté**

Dans le cadre de ce TP, il nous a été proposé d’effectuer une analyse sur deux jeux de données. Dans le premier jeu de données, nous avons différentes mesures physiologiques et médicales chez des femmes amérindiennes du peuple Pima et l’objectif est de savoir si on peut utiliser ces mesures pour modéliser voire prédire la présence/absence de diabètes chez ces femmes. Dans le deuxième jeu de données, nous avons différentes mesures et caractéristiques concernant des matériaux supraconducteurs et le but est de savoir si on peut modéliser la température critique en fonction de ces différentes mesures.

Dans les consignes, il nous a été demandé de répondre à la problématique d’un seul jeu de données et j’ai choisi d’étudier la présence ou l’absence de diabète chez les femmes du peuple Pima.

1. **Etude du jeu de données**

Le jeu de données sur le peuple Pima contient 768 observations et 9 colonnes y compris la colonne à prédire et sont toutes numériques.



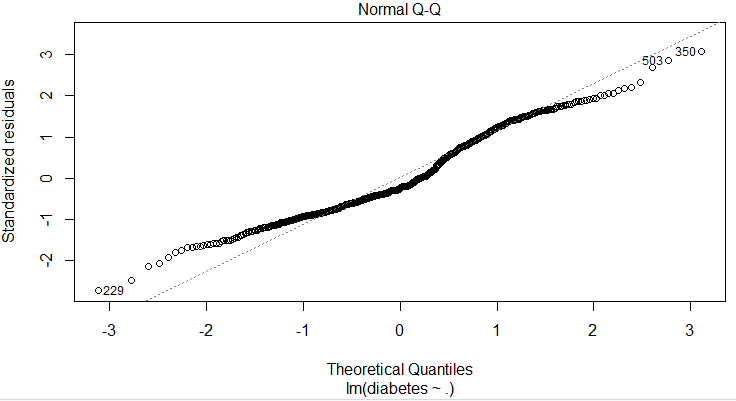
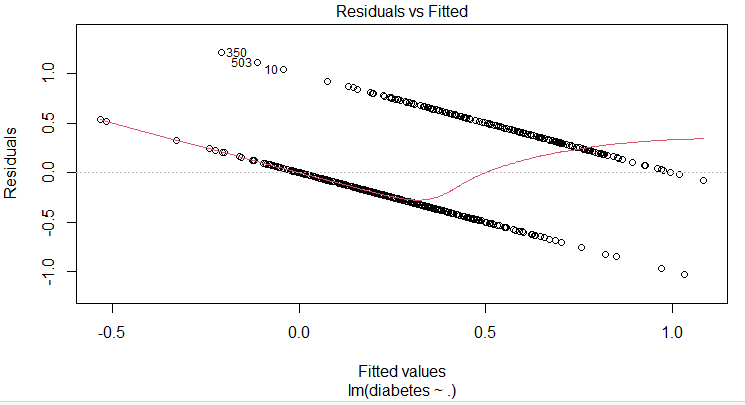
Le diagramme ci-contre montre que le jeu de données n’est pas équilibré car on dénombre bien plus de femmes saines que de femmes diabétiques. Cela pose problème car le modèle aura tendance à mieux prédire les cas sains que les cas de diabètes. L’idéal serait de d’équilibré le jeu de données en y insérant autant de cas sains que de cas de diabètes.

Les valeurs observées sont relativement petites (< 500). Dans les cas où on aurait des valeurs assez grandes (de > 103), il aurait été judicieux d’appliquer une fonction logarithmique à la colonne correspondante (log10) pour diminuer l’ordre de grandeur des valeurs et ainsi facilité les calculs. Mais ce n’est notre cas donc on utilisera les valeurs tel qu’elles sont.

Dans la suite du TP, nous allons diviser le jeu de données en 2 parties : Une partie pour entrainer le modèle (70%) et une autre pour tester le modèle (30%).

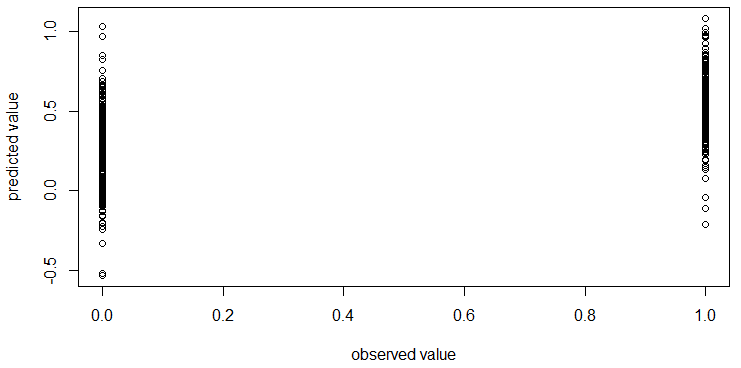
1. **Régression linéaire ou classification ?**

Dans cette partie nous allons déterminer le type d’analyse à effectuer sur ce jeu de données pour répondre à la problématique. Supposons que la régression linéaire est adapté dans ce cas. Ainsi, nous avons les résultats suivant :

C:\Users\mazou\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Capture.PNG Le R2 est très faible (R2 = 0.3015).

La moyenne des résidus n’est pas nul et la variance n’est pas constante car elle dépend fortement de la valeur prédite (**pas d’homoscédasticité**)

Les résidus **ne sont pas Gaussien** car les quantiles ne sont pas alignés sur la droite **y=x**



Les valeurs à prédire sont soit 0 ou 1 tandis que le modèle prédit des valeurs entre 0 et 1.

Au vu des résultats ci-dessus, on peut conclure que le modèle de régression linéaire n’est pas adapté à ce problème car le fait d’avoir une R2 très faible signifie que l’écart entre les valeurs prédites et les valeurs réelles est assez grand et de plus le modèle doit prédire soit 0 ou 1 et non des valeurs comprises entre 0 et 1.

Dans la suite de ce TP, nous allons donc appliquer un modèle de régression logistique (modèle de classification) pour répondre à la problématique de ce jeu de données.