TP2 : Apprentissage supervisé (Régression)

Régression simple

- On voit que la régression suit plutôt bien les données sauf au début.
- On voit que les résidus sont centrés.
- On voit que les valeurs prédites correspondent bien aux valeurs observées sauf quelles sont inversées.

Régression linéaire [SIMPLE]

- La **p-value** (Pr(>|t|)) est bien inférieure à 5% (`< 2.2e-16`). Donc on rejette (A0) car `p-value` $\leq \alpha$, $< 2.2e - 16 < 0.05$.
- La **t-value** (student test) vaut `-24.53` et **qt** (*alphaT*) vaut `1.964682`. Donc on rejette (A0) car $|-24.53| > 1.964682$.
- La **statistique de Fisher** (F-statistic) vaut `601.6` et le **qf** (*alphaFisher*) vaut `5.054041`. Donc on rejette (A0) car $601.6 > 5.054041$
- L'intervalle de confiance est `[-1.026148 ; -0.8739505]`. Donc on rejette (A0) car `0` n'est pas dans l'intervalle. `0` car la courbe peut pas être à la fois ver la gauche et vers la droite.
- Le R^2 vaut `0.5441` et le R^2 ajusté vaut `0.5432`, l'adéquation du modèle aux données est donc bien car a mis chemin entre `0` et `1`.
- Pour `10` on voit que l'intervalle de confiance est très resserré $[24.47413; 25.63256]$ et que l'intervalle de prédiction est bien plus large $[12.82763; 37.27907]$.
- Avec le test de normalité (`Shapiro-Wilk`), on obtient bien la même `p-value` (`< 2.2e-16`). `W = 0.87857` ???
- La validation croisée, un moindre carré (MSE : la moyenne des résidus au carré), nous donne `38.8901`.

Régression non linéaire : Cas polynomial

- La **p-value** : on rejette (A0) car $< 2.2e - 16 < 0.05$.
- La **t-value** : On a plusieurs β (e.g. `poly(x1, degpoly)1`, `poly(x1, degpoly)2`) on peut peut faire le test sur chacun des β pour dire s'ils sont significatifs ou non.
  - on rejette (A0) pour β_1 car $|-27.60| > 1.964682$.
  - on rejette (A0) pour β_2 car $|11.63| > 1.964682$.
- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $448.5 > 5.054041$
- L'intervalle de confiance :
  - on rejette (A0) pour β_1 car `0` n'est pas dans l'intervalle `[-163.31194 ; -141.60716]`.
  - on rejette (A0) pour β_2 car `0` n'est pas dans l'intervalle `[53.37485 ; 75.07963]`.
- Le R^2 vos `0.6407` et le R^2 ajusté vos `0.6393`.
- Avec le test de normalité (`Shapiro-Wilk`), on obtient une `p-value` plus élevée mais toujours en dessous de 5% (`6.101e-14`). `W = 0.93583` ???
- La validation croisée (MSE) nous donne `30.73622`.

Régression non linéaire : Cas spline

- La **p-value** : on rejette (A0) car $< 2.2e - 16 < 0.05$.

- La **t-value** :
  - on rejette (A0) pour β_1 car $|-19.61| > 1.964682$.
  - on rejette (A0) pour β_2 car $|-19.61| > 1.964682$.
  - on rejette (A0) pour β_3 car $|-18.78| > 1.964682$.
  - on rejette (A0) pour β_3 car $|-11.32| > 1.964682$.
- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $269 > 5.054041$
- L'intervalle de confiance :
  - on rejette (A0) pour β_1 car `0` n'est pas dans l'intervalle `[-27.60653 ; -22.57836]`.
  - on rejette (A0) pour β_2 car `0` n'est pas dans l'intervalle `[-31.08023 ; -25.42030]`.
  - on rejette (A0) pour β_3 car `0` n'est pas dans l'intervalle `[-64.08628 ; -51.94713]`.
  - on rejette (A0) pour β_4 car `0` n'est pas dans l'intervalle `[-27.62415 ; -19.45475]`.
- Le R^2 vos `0.6823` et le R^2 ajusté vos `0.6797`.
- Avec le test de normalité (`Shapiro-Wilk`), on obtient une `p-value` plus élevée mais toujours en dessous de 5% (`1.413e-15`). `W = 0.92153` ???
- La validation croisée (MSE) nous donne `27.39948`.

Régression non linéaire : Cas smoothing spline

- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $269 > 5.054041$
- Avec le test de normalité (`Shapiro-Wilk`), on obtient une `p-value` plus élevée mais toujours en dessous de 5% (`1.02e-14`). `W = 0.92929` ???
- La validation croisée (MSE) nous donne `27.95281`.

Comparaison (Régression simple)

La régression linéaire *simple* a la statistique de Fisher la plus élevée et la validation croisée la plus élevée. Ensuite, on trouve les régressions non linéaires : le `cas polynomial` qui est meilleur mais on voit avec le test de normalité (`Shapiro-Wilk`) que la `p-value` est plus élevée mais toujours en dessous de `5%`. Le `cas spline` et le `cas smoothing spline` sont similaires mais, dans notre cas, le `cas spline` et le meilleur car sa validation croisée et sa statistique de Fisher sont meilleurs. Ces deux cas ont une `p-value` plus élevée que la régression linéaire *simple* mais plus faible que le `cas polynomial`.

Régression multiple

Régression seulement sur une variable (x1)

- La **p-value** : on rejette (A0) car $< 2.2e-16 < 0.05$.
- La **t-value** : on rejette (A0) pour β_1 car $|-24.53| > 1.964682$.
- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $601.6 > 5.054041$
- L'intervalle de confiance : on rejette (A0) pour β_1 car `0` n'est pas dans l'intervalle `[-1.026148 ; -0.8739505]`.
- Le R^2 vaut `0.5441` et le R^2 ajusté vaut `0.5432`.
- Avec le test de normalité (`Shapiro-Wilk`), on obtient la même `p-value` (`< 2.2e-16`). `W = 0.87857` ???

Régression sur 2 variables

- La **p-value** : on rejette (A0) car $< 2.2e-16 < 0.05$.

- La **t-value** :
  - on rejette (A0) pour β_1 car $|-21.416| > 1.964691$.
  - on rejette (A0) pour β_2 car $|2.826| > 1.964691$.
- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $309 > 3.716066$
- L'intervalle de confiance :
  - on rejette (A0) pour β_1 car `0` n'est pas dans l'intervalle `[-1.12674848 ; -0.93738865]`.
  - on rejette (A0) pour β_2 car `0` n'est pas dans l'intervalle `[0.01052507 ; 0.05856361]`.
- Le R^2 vaut `0.5513` et le R^2 ajusté vaut `0.5495`.
- Avec le test de normalité (`Shapiro-Wilk`) on obtient la même `p-value` (`2.2e-16`). `W = 0.88947` ???
- Pour le test `anova` (ANalysis Of VAriance) on a :
  - Pour `x1` : `7.984` qui est bien supérieur au `alphaFisher` (`3.716066`)
  - Pour `x2` : `458.66` qui est bien supérieur au `alphaFisher` (`3.716066`)

Régression sur toutes variables

Pour la sélection des variables on obtient le même résultat avec pour la méthode "backward" et "forward", avec le calcul de la t-value et avec l'intervalle de confiance. On garde donc toutes les variables sauf la `2`ème (`x2`) et la `5`ème (`x5`).

> avec les 13 variables

- La **p-value** : on rejette (A0) car $< 2.2e-16 < 0.05$.
- La **t-value** :
  - on rejette (A0) pour β_1 car $|-10.347| > 1.964797$.
  - on ne rejette pas (A0) pour β_2 car $|0.052| > 1.964797$.
  - on rejette (A0) pour β_3 car $|-3.287| > 1.964797$.
  - on rejette (A0) pour β_4 car $|3.382| > 1.964797$.
  - on ne rejette pas (A0) pour β_5 car $|0.334| > 1.964797$.
  - on rejette (A0) pour β_6 car $|3.118| > 1.964797$.
  - on rejette (A0) pour β_7 car $|-4.651| > 1.964797$.
  - on rejette (A0) pour β_8 car $|9.116| > 1.964797$.
  - on rejette (A0) pour β_9 car $|-7.398| > 1.964797$.
  - on rejette (A0) pour β_10 car $|4.613| > 1.964797$.
  - on rejette (A0) pour β_11 car $|-3.280| > 1.964797$.
  - on rejette (A0) pour β_12 car $|-7.283| > 1.964797$.
  - on rejette (A0) pour β_13 car $|3.467| > 1.964797$.
- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $108.1 > 1.929413$
- L'intervalle de confiance :
  - on rejette (A0) pour β_1 car `0` n'est pas dans l'intervalle `[-0.624403622 ; -0.425113133]`.
  - on ne rejette pas (A0) pour β_2 car `0` est dans l'intervalle `[-0.025262320 ; 0.026646769]`.
  - on rejette (A0) pour β_3 car `0` n'est pas dans l'intervalle `[-0.172584412 ; -0.043438304]`.
  - on rejette (A0) pour β_4 car `0` n'est pas dans l'intervalle `[0.019448778 ; 0.073392139]`.
  - on ne rejette pas (A0) pour β_5 car `0` est dans l'intervalle `[-0.100267941 ; 0.141385193]`.
  - on rejette (A0) pour β_6 car `0` n'est pas dans l'intervalle `[0.993904193 ; 4.379563446]`.
  - on rejette (A0) pour β_7 car `0` n'est pas dans l'intervalle `[-25.271633564 ; -10.261588893]`.
  - on rejette (A0) pour β_8 car `0` n'est pas dans l'intervalle `[2.988726773 ; 4.631003640]`.
  - on rejette (A0) pour β_9 car `0` n'est pas dans l'intervalle `[-1.867454981 ; -1.083678710]`.
  - on rejette (A0) pour β_10 car `0` n'est pas dans l'intervalle `[0.175692169 ; 0.436406789]`.

    - on rejette (A0) pour $\beta_1 1$ car `0` n'est pas dans l'intervalle `[-0.019723286 ; -0.004945902]`.
    - on rejette (A0) pour $\beta_1 2$ car `0` n'est pas dans l'intervalle `[-1.209795296 ; -0.695699168]`.
    - on rejette (A0) pour $\beta_1 3$ car `0` n'est pas dans l'intervalle `[0.004034306 ; 0.014589060]`.
- Le R^2 vaut `0.7406` et le R^2 ajusté vaut `0.7338`.
- Avec le test de normalité (`Shapiro-Wilk`), on obtient la même `p-value` (`< 2.2e-16`). `W = 0.90138` ???

> avec les 11 variables

- La **p-value** : on rejette (A0) car $< 2.2e-16 < 0.05$.
- La **t-value** :
    - on rejette (A0) pour β_1 car $|-11.019| > 1.964778$.
    - on rejette (A0) pour β_8 car $|9.356| > 1.964778$.
    - on rejette (A0) pour $\beta_1 2$ car $|-7.334| > 1.964778$.
    - on rejette (A0) pour β_9 car $|-8.037| > 1.964778$.
    - on rejette (A0) pour β_7 car $|-4.915| > 1.964778$.
    - on rejette (A0) pour β_6 car $|3.183| > 1.964778$.
    - on rejette (A0) pour $\beta_1 3$ car $|3.475| > 1.964778$.
    - on rejette (A0) pour β_4 car $|3.390| > 1.964778$.
    - on rejette (A0) pour β_3 car $|-3.307| > 1.964778$.
    - on rejette (A0) pour $\beta_1 0$ car $|4.726| > 1.964778$.
    - on rejette (A0) pour $\beta_1 1$ car $|-3.493| > 1.964778$.
- La **statistique de Fisher** (F-statistic) : on rejette (A0) car $128.2 > 2.018912$
- L'intervalle de confiance :
    - on rejette (A0) pour β_1 car `0` n'est pas dans l'intervalle `[-0.615731781 ; -0.42937513]`.
    - on rejette (A0) pour β_8 car `0` n'est pas dans l'intervalle `[3.003258393 ; 4.59989929]`.
    - on rejette (A0) pour $\beta_1 2$ car `0` n'est pas dans l'intervalle `[-1.200109823 ; -0.69293932]`.
    - on rejette (A0) pour β_9 car `0` n'est pas dans l'intervalle `[-1.857631161 ; -1.12779176]`.
    - on rejette (A0) pour β_7 car `0` n'est pas dans l'intervalle `[-24.321990312 ; -10.43005655]`.
    - on rejette (A0) pour β_6 car `0` n'est pas dans l'intervalle `[1.040324913 ; 4.39710769]`.
    - on rejette (A0) pour $\beta_1 3$ car `0` n'est pas dans l'intervalle `[0.004037216 ; 0.01454447]`.
    - on rejette (A0) pour β_4 car `0` n'est pas dans l'intervalle `[0.019275889 ; 0.07241397]`.
    - on rejette (A0) pour β_3 car `0` n'est pas dans l'intervalle `[-0.172817670 ; -0.04400902]`.
    - on rejette (A0) pour $\beta_1 0$ car `0` n'est pas dans l'intervalle `[0.175037411 ; 0.42417950]`.
    - on rejette (A0) pour $\beta_1 1$ car `0` n'est pas dans l'intervalle `[-0.018403857 ; -0.00515209]`.
- Le R^2 vaut `0.7406` et le R^2 ajusté vaut `0.7348`.
- Avec le test de normalité (`Shapiro-Wilk`), on obtient la même `p-value` (`< 2.2e-16`). `W = 0.90131` ???

Comparaison (Régression multiple)

On voit que la régression sur une seule variable est moins bonne que la régression sur plusieurs, grâce à la statistique de Fisher et au R^2. Quand on utilise seulement `11` variables on voit que les intervalles de confiance sont plus restreints (donc meilleurs).

Avec le ACP

On voit grâce au ACP qu'il faut garder `5` variables, composantes, pour avoir `80%` de variance expliquée.

```
TRAINING: % variance explained
   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X    47.13    58.15    67.71    74.31    80.73    85.79    89.91    92.95
y    37.42    45.59    63.59    64.78    69.70    70.05    70.05    70.56
   9 comps  10 comps  11 comps  12 comps  13 comps
X    95.08     96.78     98.21     99.51    100.00
y    70.57     70.89     71.30     73.21     74.06
```