IUT de Paris - Rives de Seine
Université Paris Cité

SAE « Régression sur données réelles »

M. Bou Aziz

République tchèque - Maroc

William Lefebvre
Célia Zuk

BUT SD1
Groupe 14

Table des matières

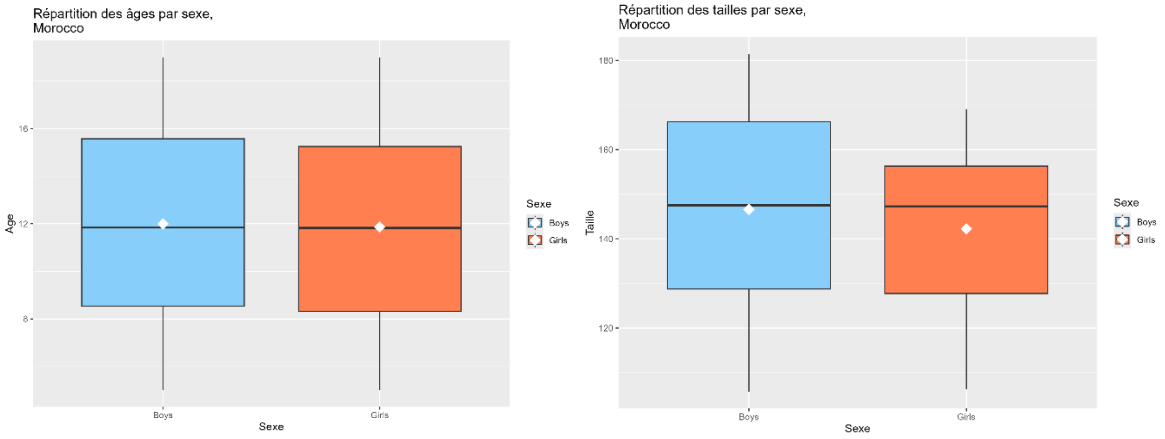
Introduction

En 2020 une étude a été menée dans près de 90 pays pour obtenir des informations sur les courbes de croissances de jeunes de 5 à 19ans. Dans chaque pays, des échantillons de 2000 filles et 2000 garçons ont été tirés aléatoirement et leur taille a été mesurée en centimètres.
Dans ce document, nous étudierons deux pays de cette étude, le Maroc et la République Tchèque. Avec des analyses statistiques, des ajustements et à l'aide de régressions, nous répondrons à la problématique suivante :

Quelles observations pouvons-nous réaliser à partir de nos données ?

Pour répondre à cette problématique, nous allons commencer par procéder pays par pays. Dans deux premières parties, pour le Maroc puis pour la République tchèque, nous décrirons statistiquement les données à disposition puis étudierons la croissance des filles puis celle des garçons, après quoi nous comparerons la croissance des sexes.
Dans une troisième partie nous étudierons les courbes de régressions entre les deux pays.
Nous conclurons ensuite en résumant nos observations puis effectuerons un résumé en anglais comme demandé dans les consignes de cette SAE.

Maroc



Sur ce premier graphique est représenté la distribution de l'âge en fonction du sexe. Nous pouvons remarquer que la répartition des âges entre les filles et les garçons est quasiment identique, allant de 5 ans à 19 ans. En ce qu'il concerne les quartiles, le premier est de 8,5 ans pour les garçons et 8.3 ans pour les filles, ce qui représente que 25% des garçons et des filles répertoriés ont un âge inférieur à 8.5 ans, pour le troisième quartile, il est de 15 ans pour les deux genres cela signifie que 75% des filles et des garçons ont un âge inférieur à 15 ans, la médiane quant à elle est de 11.8 ans. Enfin, la répartition des valeurs de l'âge des filles et des garçons est égale, l'échantillon a été construit avec une répartition similaire des âges pour les deux genres.
Ce second graphique, représente la distribution des tailles (en cm) pour les garçons et les filles au Maroc.
Concernant la boite à moustache représentant les garçons, la médiane est d'environ 147.5 cm, le premier quartile est de 128 cm cela veut dire que 25% des tailles sont inférieur à 128 cm. Le troisième quartile quant à lui, est de 166 cm. Pour la dispersion des données, elles vont de 105 cm à 181 cm. Le symbole en forme de losange représente la moyenne, nous pouvons constater qu'elle est de 146 cm.
Pour la boite à moustache représentant les filles, la médiane est d'environ 147 cm comme celle des garçons. Concernant les quartiles, le premier est de 127 cm, ce qui signifie que 25% des filles mesurent moins de 130cm et le troisième de 156 cm, indiquant que 75% des filles mesurent moins de 156cm. La dispersion des données s'étant de 106 cm à 169 cm. Pour la moyenne, elle est de 142 cm.
Enfin, les tailles médianes des garçons et des filles sont similaires, elles sont de 147 cm. Néanmoins, nous remarquons une plus grande dispersion des tailles pour les garçons que pour les filles.

Filles

Nous regardons dans un premier temps l'évolution de la courbe concernant les jeunes marocaines.

Ajustements

Nous allons tenter d'ajuster cette évolution de différentes manières. Nous allons réaliser un ajustement linéaire, puis des ajustements de degré 1, 2, 3 et 4.

Pour plus d'information sur la méthode utilisée tout au long du document pour les ajustements et les R^2 se référer à l'annexe.

Pour la croissance des jeunes marocaines, l'expression de la droite des moindres carrés sera :

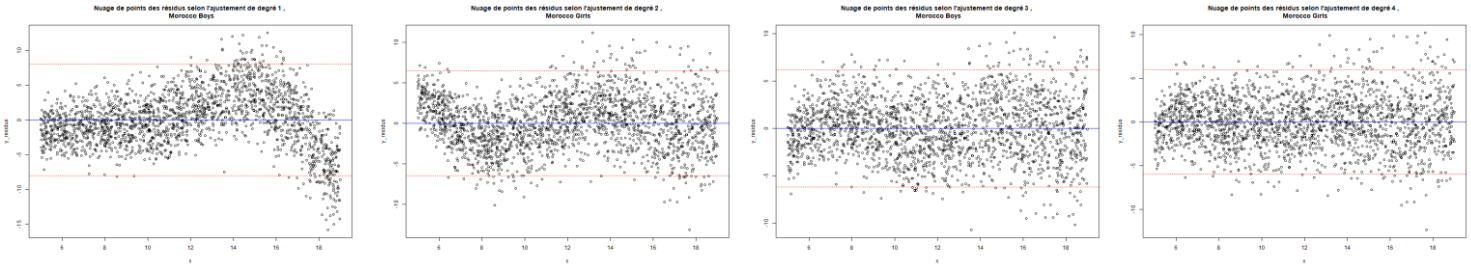
$$y_1 = 96.98215 + 3.812552\,x$$

La valeur du R^2 est de 0.8894131
Environ 88,94% de la variance de la taille est expliquée par notre ajustement linéaire.

Nous pouvons donner les expressions obtenues pour les 3 autres ajustements :

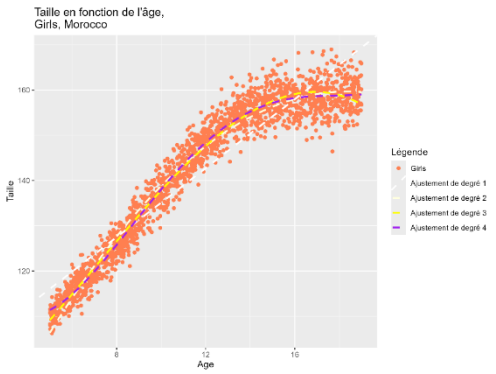
$$y_2 = 58.89509 + 10.98194x - 0.2991582x^2$$
$$y_3 = 86.62061 + 2.883076x + 0.4244457x^2 - 0.02009224x^3$$
$$y_4 = 146.4807 - 20.83883x + 3.729824x^2 - 0.2133316x^3 + 0.004030457x^4$$

Nous pouvons constater que les valeurs des R^2 sont respectivement de 0.9605175, 0.964577 et 0.9665979.
Sur ces 3 ajustements, un peu plus de 96% de la variance est expliquée. Le gain est modeste pour chaque degré ajouté.



Pour appuyer les R^2, on peut regarder les graphiques des résidus des ajustements. De gauche à droite, les ajustements de degrés 1 à 4.
Dans les 3 premiers, on observe une tendance dans les résidus. L'ajustement linéaire par exemple va sous-estimer les tailles pour les marocaines de 14ans et les surestimer pour celles de 16 à 18ans. Aux degrés 2 et 3, la tendance s'estompe même si l'on continue d'apercevoir une sorte de tendance sinusoïdale.

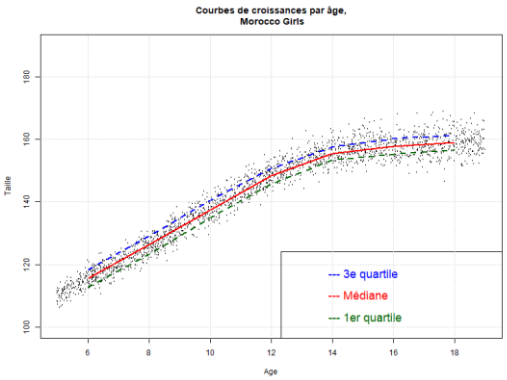
Au degré 4, la tendance s'estompe significativement. On va observer dans ces résidus, pour des âges à partir de 12 ans, des valeurs qui s'éloignent de plus de deux écarts-type (traits rouges) de 0 et on peut en déduire que la dispersion des tailles est plus importante à partir de cet âge.



Avec ce graphique ci-contre, on peut affirmer que l'ajustement linéaire est le moins précis. Plus le degré augmente, plus on gagne de qualité sur notre ajustement.
L'ajustement de degré 4 est celui qui explique le mieux nos données.

A partir de cet ajustement de degré 4, nous observons sur la courbe des jeunes marocaines, une **accélération de la croissance vers 7 ans** puis une **diminution aux alentours de 13.5 ans**. A 19ans, la taille estimée est de 159.02cm.

Courbe de régression



Les courbes de croissances suivent des évolutions similaires, la croissance semble uniforme au sein de la population des jeunes marocaines.
On peut noter qu'à 14ans, la courbe du 1er quartile se rapproche de celle de la médiane. Les marocaines les plus petites, ne le sont que de peu et leur taille est proche de la taille médiane.

Au moment de l'adolescence, dès 13-14 ans, on note une dispersion au niveau des tailles extrêmes. Il y a plus de disparité que dans l'enfance, même si cela est peu perceptible en regardant les 1er et 3e quartiles.

Garçons

Regardons maintenant l'évolution de la croissance chez les jeunes marocains.

Ajustements

Afin d'analyser les tailles des filles en République Tchèque, plusieurs ajustements ont été effectués allant d'un ajustement linéaire à des ajustements polynomiaux de degré 2, 3 & 4.

L'ajustement linéaire est donné par l'équation suivante :

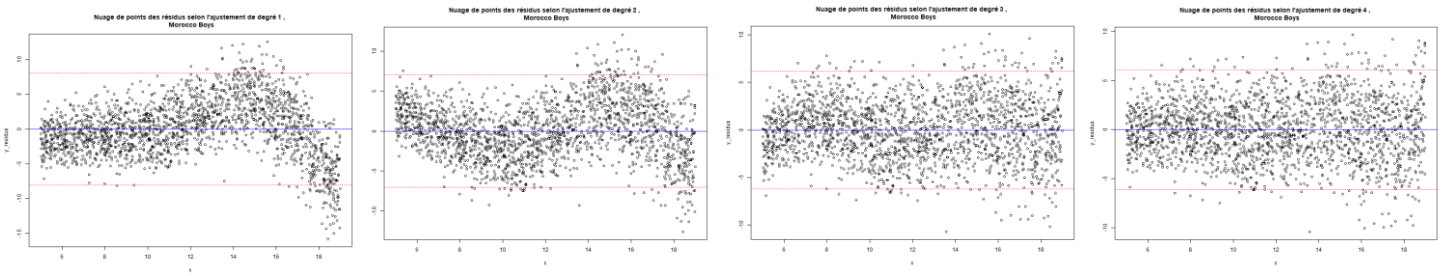
$$y_1 = 87.49605 + 4.925841 * x$$

La valeur du R^2 est de 0.9611428.
Environ 96.11% de la variance des tailles peut être expliquée par l'âge avec ce modèle linéaire.

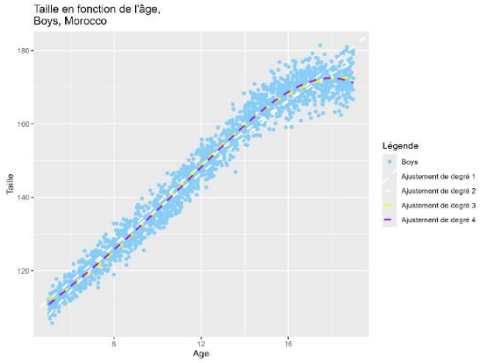
Les ajustements 2, 3, et 4, sont représentés par les équations suivantes :

$$y_2 = 70.8576 + 8.046389 * x - 0.1297423 * x^2$$
$$y_3 = 114.8194 - 4.852735 * x + 1.02609 * x^2 - 0.03215059 * x^3$$
$$y_4 = 76.91003 + 10.18465 * x - 1.066924 * x^2 + 0.08996644 * x^3 - 0.002541195 * x^4$$

Les valeurs des R^2 pour les ajustements précédents sont respectivement de 0.9699567, 0.9768497 et 0.97739.



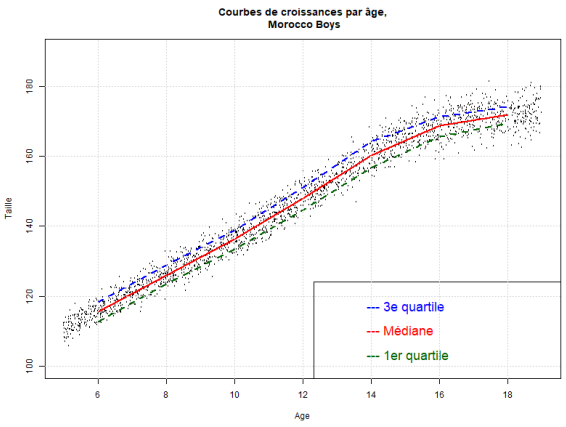
Les graphiques des résidus pour chaque ajustement montrent une répartition de plus en plus homogène autour de 0.
Au degré 1, les tailles seront sous-estimées vers 14-15ans et surestimées vers 17-19ans. Aux degrés 2 et 3, une tendance avec une forme sinusoïdale est perceptible. Cette tendance s'estompe au degré 4, même si on constate une dispersion au niveau des résidus qui reste tout au long des âges et que celle-ci est d'avantage prononcée sur les âges élevés. Ce dernier ajustement est celui des 4 qui a la meilleure capacité à expliquer la variance des tailles et à capturer les structures des données.



Pour les garçons marocains, tous les ajustements semblent juste.
Si l'ajustement de degré 4 est le plus qualitatif (en termes de R^2), il va estimer (à tort) un recul de la taille à partir de 17ans.

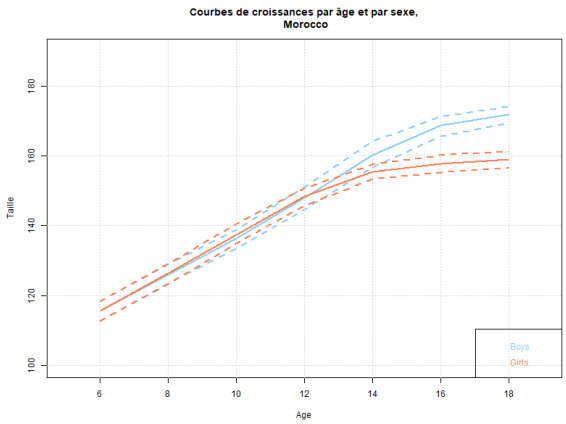
D’après cet ajustement de degré 4, la croissance est **quasi-linéaire de 5 à 16ans**. Si on se base sur l’ajustement linéaire, près de 5cm sont gagnés par ans. On observe dans cette croissance une **inflexion à partir de 16 ans**. A 19ans, la taille estimée sera de 171.17cm.

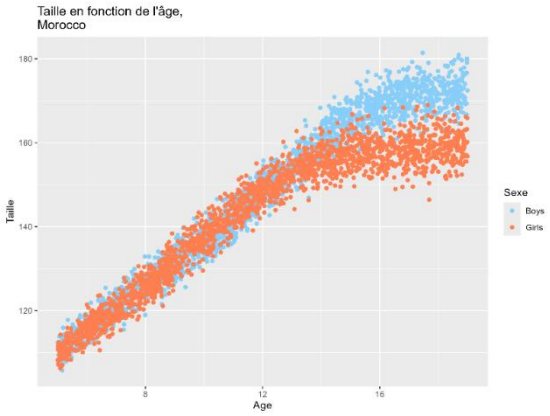
Courbe de régression



La variabilité de la croissance est constante, toutes les courbes de croissances évoluent de la même manière et l’écart interquartile reste constant.
On peut donner l’âge de 14 ans où la courbe du troisième quartile s’écarte légèrement des deux autres, une partie des jeunes marocains grandiront plus vite que les autres à cet âge.
A partir de 15 ans, on constate une plus forte dispersion au niveau des tailles extrêmes.

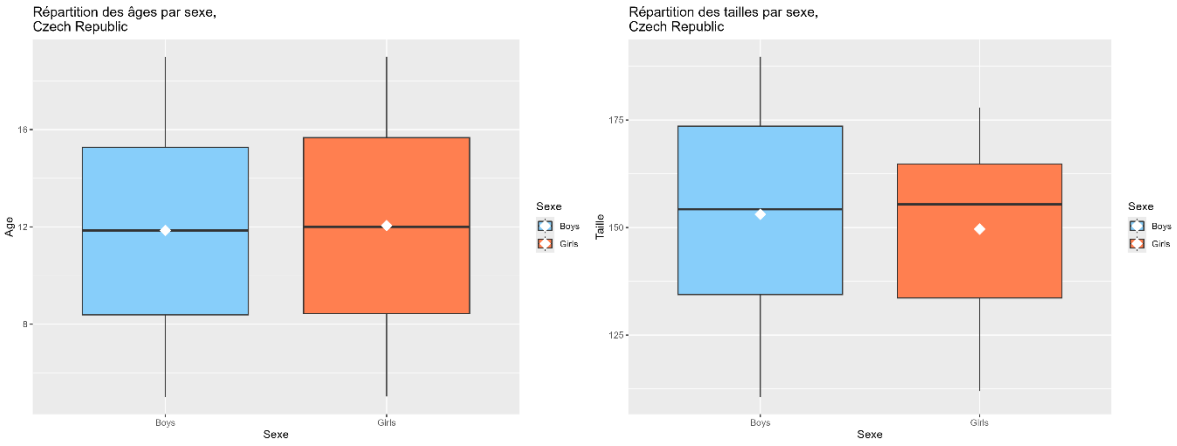
Comparaison entre les sexes





Pour comparer les tailles des jeunes marocains et marocaines. On voit sur nos données que les jeunes enfants (6-7ans) ont sensiblement la même taille. Les filles vont grandir plus tôt et plus fort, dès 8-9ans. Jusqu’à 11-12ans elles seront plus grandes que les garçons. En revanche leur croissance ralentira à partir de 12ans alors que les garçons, eux, continueront de grandir. Cela explique les écarts de taille à l’âge adulte.
Que ce soit en médiane ou en quartiles, l’écart de taille entre les sexes au début de l’âge adulte sera de près de 20cm.

République tchèque



Le boxplot de droite représente la distribution de l'âge en fonction du sexe.
Nous pouvons constater qu'encore une fois il n'y a pas de grande différence entre les garçons et les filles, néanmoins on peut observer une légère différence au niveau de la dispersion des valeurs qui est légèrement supérieure pour les filles. En ce qu'il concerne la médiane, elle est de 12 ans pour les filles, contre 11,8 ans pour les garçons, il y a une légère différence. Pour les premiers quartiles, il est de 8.3 pour les garçons et de 8.4 pour les filles. Pour le troisième quartile, il est de 15.2 pour les garçons et de 15.6 pour les filles.
Encore une fois, les valeurs de l'échantillon ont été construit avec une répartition similaire des âges pour les deux genres.

Le second graphique montre la répartition des tailles entre les garçons et les filles en République Tchèque. On peut noter que les médianes sont très proches, 154.3 cm pour les garçons et 155.4 cm pour les filles. Les moyennes sont également proches, 153.1 cm pour les garçons et 149.6 cm pour les filles. En ce qu'il concerne les quartiles, le premier est de 134.4 cm pour les garçons et de 133.6 cm pour les filles, ce qu'il veut dire que pour chaque genre, 25% d'entre eux ont une taille inférieure à ce premier quartile. Pour le troisième, on note que 75% des garçons ont une taille inférieure à 173.6 cm, et que 75% des filles ont une taille inférieure à 164.8 cm.

On observe néanmoins une dispersion des valeurs de l'échantillon plus grande pour les garçons que pour les filles, les tailles des garçons varient entre 110.6 cm et 189.7 alors que celles des filles varient de 112 cm à 177.8 cm.

Fille

Ajustements

Afin d'analyser les tailles des filles en République Tchèque, plusieurs ajustements ont été effectués allant d'un ajustement linéaire à des ajustements polynomiaux de degré 2, 3 & 4.
L'ajustement linéaire est donné par l'équation suivante :

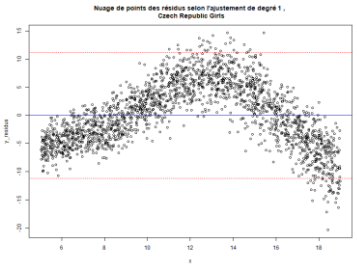
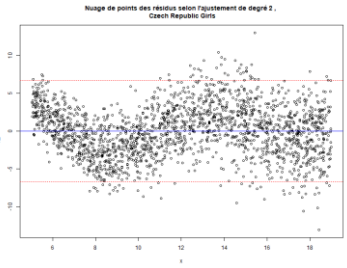
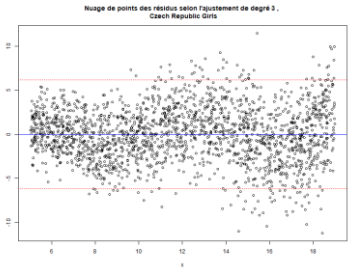
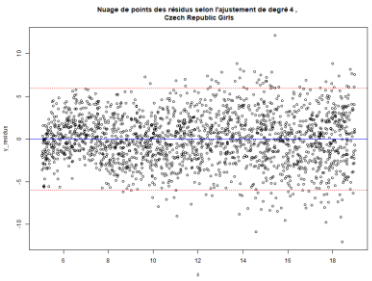
$$y_1 = 101.3631 + 4.005476 * x$$

Nous pouvons constater que la valeur du R^2 est de 0.8969578.
Environ 89.7% de la variance des tailles peut être expliquée par l'âge avec ce modèle linéaire.

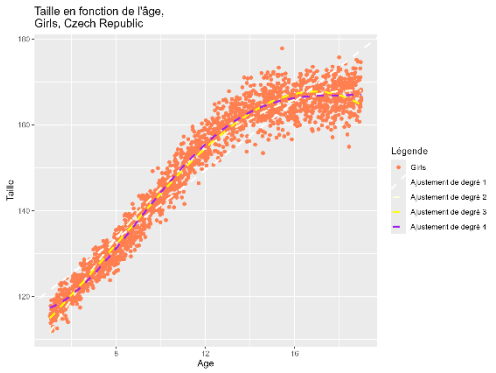
Les ajustements 2, 3, et 4, sont représentés par les équations suivantes :

$$y_2 = 62.47027 + 11.31044 * x - 0.31428 * x^2$$
$$y_3 = 96.62169 + 1.305488 * x + 0.5917303 * x^2 - 0.0248426 * x^3$$
$$y_4 = 164.8883 - 25.77672 * x + 4.364439 * x^2 - 0.245207 * x^3 + 0.004590958 * x^4$$

Les valeurs des R^2 pour les ajustements précédents sont respectivement de 0.962834, 0.96833577 et 0.9706672.
Ces ajustements montrent une amélioration progressive de la capacité à expliquer la variance des tailles avec des modèles de degré plus élevé. Le modèle de degré 2 explique 96.28% de la variance des tailles, tandis que le modèle de degré 3 en explique 96.84%, et le modèle de degré 4 atteint 97.07%.

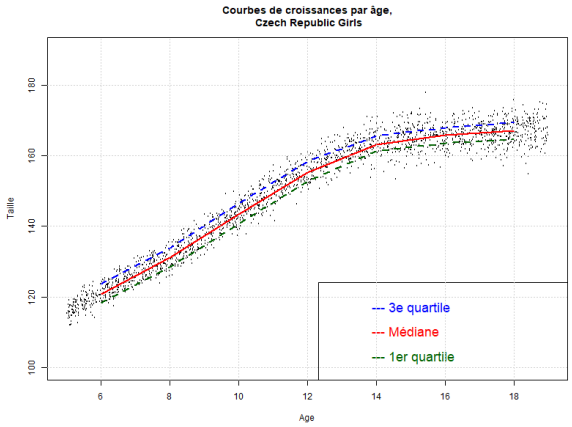





Les graphiques des résidus pour chaque ajustement montrent une répartition de plus en plus homogène autour de 0, avec le modèle de degré 4 présentant la meilleure adéquation. Les ajustements de degrés supérieurs réduisent les structures apparentes dans les résidus, indiquant une meilleure capture des variations dans les données.



L'ajustement de degré 4 est celui qui approche le mieux les données, notamment sur les âges élevés. Sur cet ajustement, on voit une **accélération de la croissance vers 6 ans**. On voit aussi une inflexion de la courbe et donc **un ralentissement de la croissance vers 14ans.**

Courbe de régression



Les courbes de croissances évoluent de manière similaire.
La courbe du 1er quartile est plus proche de la courbe de la médiane que celle du 3e quartile ne l'est. Les petites tchèques sont moins petites que les grandes ne le sont. C'est notamment le cas entre 12 et 14ans. La dispersion des tailles extrêmes est plus prononcée à partir de 16ans, mais elle l'est déjà depuis 13ans.

Garçons

Ajustements

L'ajustement linéaire est donné par l'équation suivante :

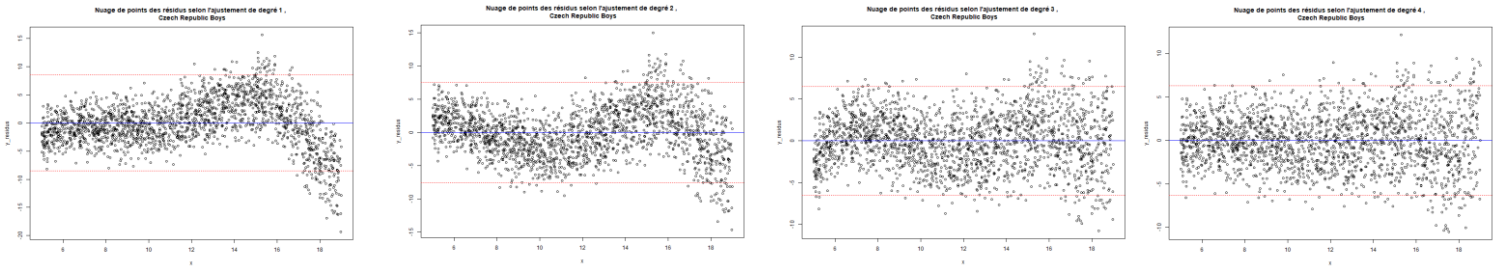
$$y = 91.69468 + 5.178827x$$

Nous pouvons constater que la valeur du R^2 est de 0.9596329.
Environ 95.96% de la variance des tailles peut être expliquée par l'âge avec ce modèle linéaire simple.

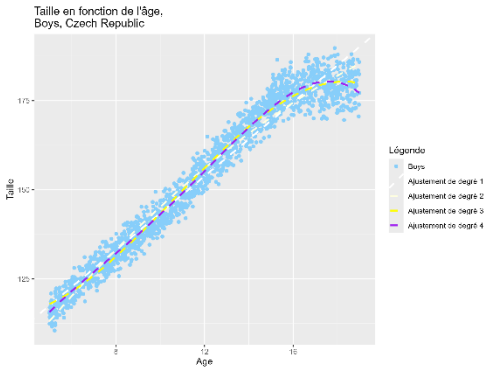
En ce qu'il concerne les ajustements de degré 2,3 et 4, on note les expressions suivantes :

$$y_2 = 74.33632 + 8.470455x - 0.1383041x^2$$
$$y_3 = 123.3888 - 5.982468x + 1162338x^2 - 0.03635704x^3$$
$$y_4 = 57.36073 + 20.42474x - 2.545517x^2 + 0.1818601x^3 - 0.004579037x^4$$

Les valeurs des R^2 pour les ajustements ci-dessus sont respectivement : 0.9685073, 0.976378 et 0.9779203.

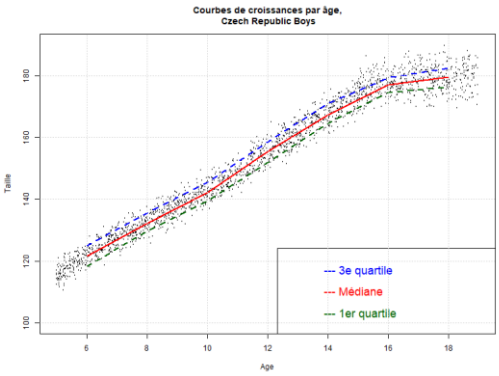


Ci-dessus, les trois graphiques de droite, montrent les résidus des modèles de régression polynomiale de degrés 2, 3 et 4 pour les garçons de la République tchèque. Les ajustements montrent une amélioration progressive de la capacité à expliquer la variance des tailles avec des modèles de degré plus élevé. Le modèle de degré 2 explique 96.85% de la variance des tailles, tandis que le modèle de degré 3 en explique 97.64%, et le modèle de degré 4 atteint 97.79%.
De plus, les graphiques des résidus pour chaque ajustement montrent une répartition de plus en plus homogène autour de 0, avec le modèle de degré 4 présentant la meilleure adéquation. En résumé, les modèles de degré supérieur offrent une meilleure capacité à expliquer la variance des tailles et à capturer les structures des données.



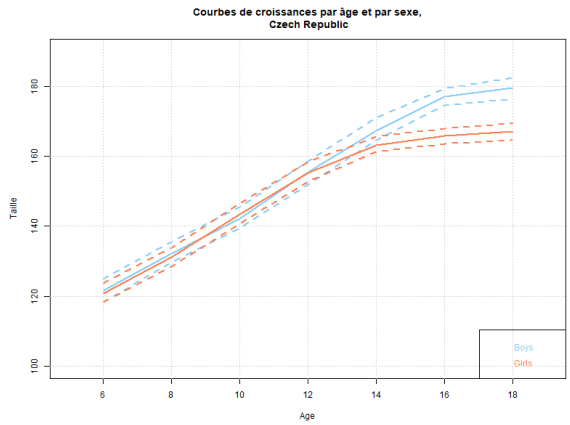
Sur ce graphique on voit que les ajustements sont justes dès le premier degré. La relation entre la taille et l'âge est relativement linéaire. La croissance est **quasi-constante de 5 à 16ans**.
En se basant sur l'ajustement de degré 4, on peut dire qu'il y a un **ralentissement de la croissance vers 16-17ans**.

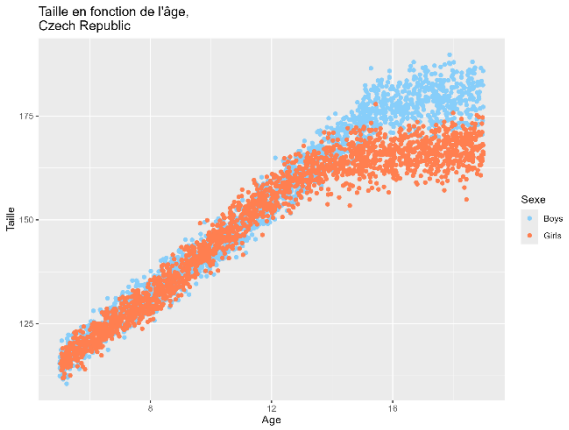
Courbe de régression



Les courbes de croissances évoluent de manières similaires. Sur chacune des courbes, la croissance de la taille augmente à l'âge de 10 ans pour ralentir vers 16 ans.
L'écart interquartile est constant tout au long des années. A partir de 15ans, des tailles extrêmes se détachent de la masse. Certains tchèques sont notablement plus petits ou plus grands que les autres.

Comparaison entre les sexes

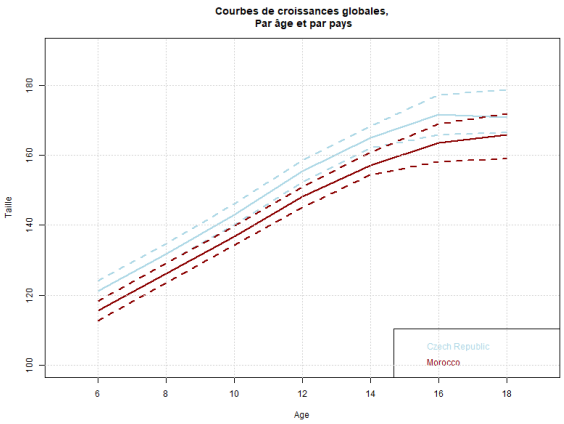




Pour comparer les tailles des jeunes tchèques, on voit sur nos données qu'entre 6 et 8 ans, un très jeune tchèque est légèrement plus grand qu'une très jeune tchèque.
De 9 à 12 ans, la taille des filles augmente plus fortement que celle des garçons. Elles seront même plus grandes pendant une brève période.
Cependant, à partir de 12ans leur croissance ralentit contrairement à celle des garçons qui ne ralentira qu'à partir de 16ans. Cela explique les écarts de tailles à l'âge adulte. Au début de celui-ci, elles seront plus petites en médiane de près de 15cm.

Comparaison entre pays

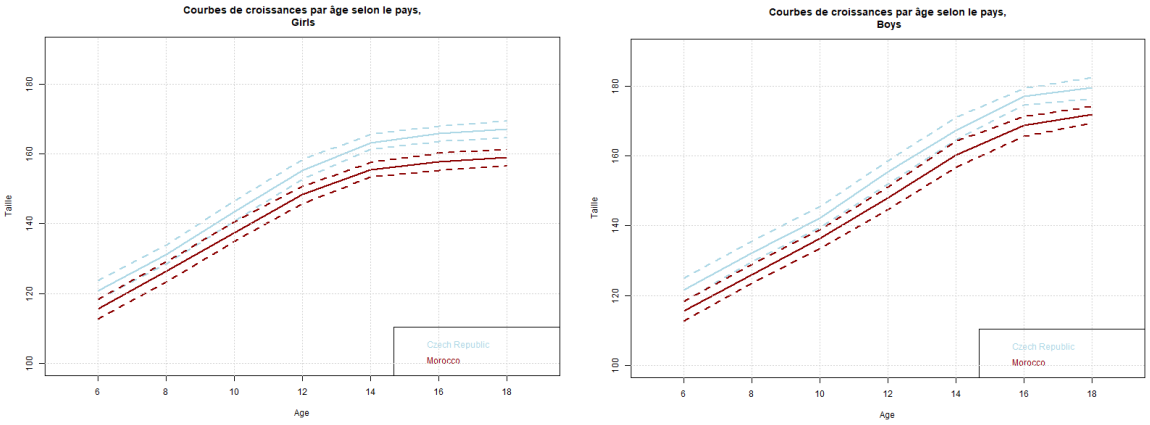
Comparaison globale



Si on compare les tailles du Maroc et de la République Tchèque indépendamment des sexes, on voit que quel que soit leur âge, **les jeunes tchèques sont bien plus grand que les jeunes marocains**.
Jusqu'à 14-15ans, 75% des plus grand tchèques sont plus grands que les 75% des plus petits marocains. Ils ont en médiane près de 5cm d'écart.
Cette différence se réduit à l'âge adulte, où ces 75% marocains ne sont dépassé en taille que par un peu moins de 50% des tchèques.

Comparaison détaillée

Nous pouvons pousser cette analyse en fonction des sexes.



En poussant l'analyse entre les pays selon les sexes, la conclusion reste la même, les tchèques sont plus grand que les marocains. Même les proportions restent sensiblement les même, nous obtenons toujours que 75% des tchèques les plus grands sont plus grands que 75% des marocains les plus petits.
Nous pouvons toutefois noter qu'au début de l'âge adulte, l'écart ne se réduit pas comme il le fait sur la population globale. Au contraire, il se creuse.

Conclusion

Dans nos données, nous observons une différence entre les sexes ainsi qu'une différence entre les pays.
Au niveau des sexes, les filles ont une plus forte croissance qui commence plus tôt mais se finie plus tôt. Celle des garçons est plus linéaire comme en témoigne les forts R^2 de leur ajustement linéaire. Si cette constante vient au prix d'un croissance plus lente, elle est compensée par le fait que leur croissance se finie plus tard. C'est la source des différences de tailles à l'âge adulte.
On observe également une différence au niveau de la région du monde. D'un pays à l'autre, les individus seront plus grands. Que ce soit durant l'enfance et l'adolescence ou à l'âge adulte et quel que soit leur sexe. En effet, si une fille d'un pays sera plus petite que celle d'un autre pays, elle aura une courbe de croissance similaire avec pour seule différence un décalage de quelques centimètres.

Concernant le projet, ce travail nous a permis de réaliser l'importance de la visualisation de données et l'utilité de l'analyse des courbes de régressions. De plus, les différents ajustements nous ont fait réaliser que nous pouvions analyser leur qualité pour émettre des hypothèses sur la nature d'une relation.
La rédaction du rapport nous a permis d'adopter une démarche méthodique dans notre analyse et dans la présentation des résultats.
Le projet nous a également permis de travailler sur des codes informatiques et de réaliser l'importance de réaliser un bon code (voir annexes). Construit autour de boucles et de noms de variables stockées dans des vecteurs, il nous a permis de réaliser des graphiques à la chaine exportés sous forme d'images ainsi que d'ajuster titres et couleurs avec aise.
Pour finir, le travail en binôme a été constructif. Travailler avec des nouvelles personnes nous a permis de nous a permis d'échanger et de progresser sur notre vision du travail en équipe. Cela nous a permis d'adopter de nouvelles méthodes de travail.

Summary

To sum up this document in English.

After a thorough analysis of our data concerning young people from two different countries, we were able to come to 2 conclusions.
The first conclusion is that girls and boys will grow up differently. On one hand, young girls will grow up earlier and faster but this growth will slow down early (around 12years old). On the other end, the growth for boys will be smoother and a little bit slower, but it will go on until 14-16 years old before slowing. The longer growth period explain the different of height between young men and young women.
The second conclusion is that height will vary depending on the geographical location. Young people from a country will be taller or smaller. This observation doesn't seems to be affected by the age of the teenagers, as a "taller population" will be taller whatever the age we are looking at.
Moreover, the difference between country only seems to affect the height, girls and boys will have a similar growth curve that will only be shifted upward (or downward) by a couple of centimeters.

Annexes

Méthodologie ajustements et R^2

Nous donnerons ici la méthode de l'ajustement linéaire. Dans cette méthode, nous recherchons la droite des moindres carrés, $y = ax + b$
C'est la droite qui va minimiser la somme des écarts entre les valeurs prédites par la droite et les valeurs observées.
Pour ce faire on sait les valeurs que vont prendre a et b :

$$a = cov(x, y)/var(x) \text{ et } b = mean(y) - a * mean(x)$$

Ici, ayant accès au logiciel R, nous allons utiliser la commande `lm_model<-lm(y~x)` et, à l'aide d'une commande `summary()` nous allons en extraire les coefficients. Nous ferons de même pour les coefficients des autres ajustements en adaptant la formule.

A partir d'une formule d'un ajustement nous pouvons réaliser des estimations sur la taille à partir des âges. Nous pouvons mesurer la qualité de cet ajustement à l'aide d'un coefficient de détermination R^2.
Rappelons que nous pouvons calculer ce coefficient, sur les ajustements linéaires, en calculant $cor(x, y)^2$. Dans ce document, à nouveau, nous l'obtiendrons à l'aide de la commande `summary()` sur R.

Scores AIC

Sorties logiciel affichant les scores AIC des différents ajustements réalisés

```
Morocco Girls : les différents AIC (par ordre croissant de degré d'ajustement
) :
 12442.25 10384.36 10169.37 10053.88
```

```
Morocco Boys : les différents AIC (par ordre croissant de degré d'ajustement)
:
 11235.41 10722.9 10203.64 10158.41
```

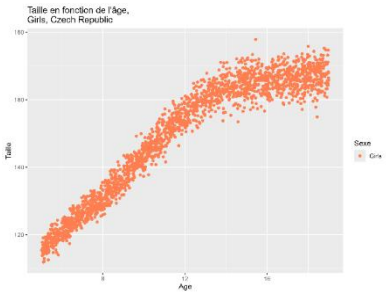
```
Czech Republic Girls : les différents AIC (par ordre croissant de degré d'aju
stement) :
 12556.08 10518.59 10198.79 10049.21
```

```
Czech Republic Boys : les différents AIC (par ordre croissant de degré d'ajus
tement) :
 11480.4 10985.89 10412.73 10279.69
```

Le degré d'ajustement avec le plus faible score AIC correspond au « meilleur » ajustement, celui qui offre le meilleur compromis entre précision et complexité.

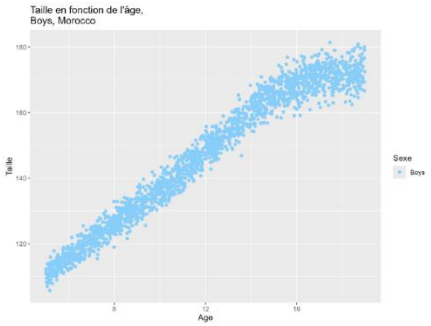
Premières impressions sur les relations âge/taille

Maroc-filles



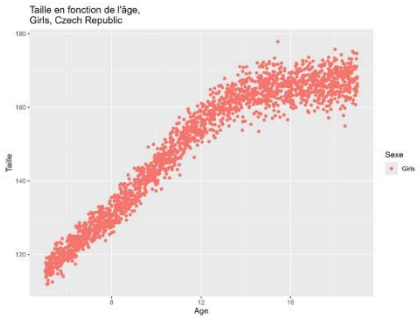
Visuellement, de 5 à 12.5 ans les jeunes marocaines semblent grandir de manière uniforme et linéaire, elles commencent à environ 115cm à 5 ans pour arriver à un peu plus de 160cm à 12.5 ans.
A partir de ces 12.5 ans, leur croissance ralentie fortement.

Maroc-garçons



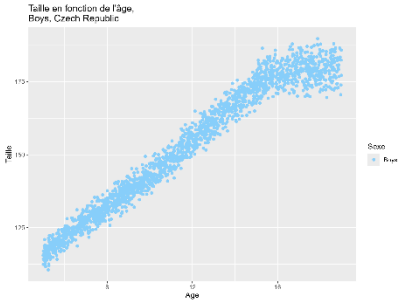
A première vue, les marocains semblent croître de manière constante de 5 à 16 ans, âge auquel ils mesureront un peu moins de 170cm. Sur les deux années qui suivent, leur croissance sera très ralentie.

République tchèque-filles



Ce graphique représente la croissance en taille des filles en République Tchèque. On observe une croissance rapide pendant l'enfance puis une stabilisation pendant l'adolescence, à partir de 14 ans.

République tchèque-garçons



Ce nuage de point représente la croissance en taille des garçons en République Tchèque. On observe une croissance constante jusqu'à 15-16ans puis une stabilisation ou du moins un ralentissement après cet âge.

Extraits du code R

Fonction réalisant des nuages de points pour les résidus

```
ndp_des_residus <- function (x, y_residus, pays, sexe, degre) #réalise un ndp des résidus
{
 p <- plot(x,y_residus, main=paste("Nuage de points des résidus selon l'ajustement de degré",degre,", \n", pays, sexe))
 abline(0,0,col="blue")
 abline(2*sd(y_residus), 0, col="red", lty=3)
 abline(-2*sd(y_residus), 0, col="red", lty=3)
 # legend("bottomleft", legend=c("--- +/- deux écart-type"),
 #       text.col=c("red"), cex=1.5)
 p_combiné <- recordPlot()
 save_plot(monplot = p_combiné, nom_image = paste0(pays, sexe, "-ajustement D", degre, "-residus ndp"))
}
```

Une des fonctions permettant l'export des graphiques

```
save_ggplot <- function(monplot, nom_image, force_export = 0)
{
 if (exporter == 0 & force_export == 0){return (0)}
 timecode <- format(Sys.time(), "%m-%d-%Hh%M")
 if (appareil == "Home ONE")
 {
   savepath = "C:/Users/217wi/OneDrive - Université de Paris/Cours BUTSD1/S2/Projet/Regression/SAE/Plots"
 }
 ggsave(file.path(savepath, paste0(timecode, nom_image, ".png")), plot = monplot, width = 8, height = 6, dpi = 300)
}
```

Boucle permettant de réaliser les ajustements et courbes de régressions par pays et par sexe (extraits)

```
#Par pays, par sexe -> ajustement polynomial
for (pays in l_pays)
```

```
{
  for (sexe in l_sexe)
  {
    temp <- donnees %>% filter(Pays == pays) %>% filter(Sexe == sexe)
    g<-ggplot (data = temp) + aes_string(x="Age", y="Taille", color="Sexe") + geom_point()
+ ggtitle(paste0("Taille en fonction de l'âge, \n", sexe, ", ", pays)) +
scale_color_manual(values = vect_couleurs)
    print(g)
    save_ggplot(monplot=g, nom_image = paste0(pays,sexe, "-ndp taille-age"))

    x <- temp$Age
    y <- temp$Taille

[...]

    #POLYNOMIAL 4
    model_poly4<-lm(y~x+I(x^2)+I(x^3)+I(x^4))
    s<-summary(model_poly4)
    cat("Ajustement de degré4 :", s$coefficients[1,1], "+", s$coefficients[2,1],"x +",
s$coefficients[3,1], "x^2",s$coefficients[4,1], "x^3", s$coefficients[5,1], "x^4", "(",pays,
sexe,")")
    cat("La valeur du R2 pour l'ajustement de degré 4 est de : \n", s$r.squared, "\ncad que
x% de la variance est expliquée par ce modèle. C'est bien/pas bien. (", pays,
sexe,")\n\n")

    g_ajustements <- g_ajustements +
geom_smooth(method=lm,formula=y~x+x+I(x^2)+I(x^3)+I(x^4),se=FALSE,
aes(color="Ajustement de degré 4"), linetype="dashed") + labs(color = "Légende")
    print(g_ajustements)
    save_ggplot(monplot=g_ajustements, nom_image = paste0(pays,sexe, "-ndp taille-
age ajustement d4"))

    y_fitted <- fitted(model_poly4)
    y_residus <- y - y_fitted
    ndp_des_residus(x = x,y_residus = y_residus, pays = pays, sexe = sexe, degre=4)

  #affichage des AIC pour interprétation
  cat(pays, sexe, ": les différents AIC (par ordre croissant de degré d'ajustement) : \n",
AIC(lm_model), AIC(model_poly2), AIC(model_poly3), AIC(model_poly4), end = "\n")

  #COURBE DE REGRESSION
```

```
[...]
  newx<-cut(x,breaks=c(5,7,9,11,13,15,17,19))
  # levels(newx) #les modalites de cette nouvelle variable
  # table(newx) #on verifie qu'il y a suffisament d'observations dans chaque classe de
donnees -> le découpage est ok, 260-330 enfants par classe
  newy_median<-tapply(y,newx,median) #mediane
  newy_q1<-tapply(y,newx,q1_maison)
  newy_q3<-tapply(y,newx,q3_maison)
  #plot courbe de croissance médiane/q1/q3
  p<-plot(x,y,main=paste("Courbes de croissances par âge,\n", pays, sexe),
        xlim = c(5,19),
        ylim = c(100, 190),
        xlab="Age",ylab="Taille", cex=0.1) #ajouter (retirer) le type="n" si on veux (pas) les
points
  lines(c(6,8,10,12,14,16, 18),c(newy_median),type="l",col="red", lwd=2)
  lines(c(6,8,10,12,14,16, 18),c(newy_q1),type="l",col="darkgreen", lwd=2, lty=2)
  lines(c(6,8,10,12,14,16, 18),c(newy_q3),type="l",col="blue", lwd=2, lty=2)
  legend("bottomright", legend=c("--- 3e quartile", "--- Médiane", "--- 1er quartile"),
       text.col=c("blue","red","darkgreen"), cex=1.5)
  grid()
  p_combine <- recordPlot()
  save_plot(monplot = p_combine, force_export = 0, nom_image = paste0(pays, sexe, "-
courbes de croissances"))
 }
}
```

Exemples de sorties générées par le code

```
Ajustement linéaire : 101.3631 + 4.005476 x ( Czech Republic Girls )
Ajustement de degré2 : 62.47027 + 11.31044 x + -0.3031428 x^2 ( Czech Republi
c Girls )
Ajustement de degré3 : 96.62169 + 1.305488 x + 0.5917303 x^2 -0.02484726 x^3
( Czech Republic Girls )
Ajustement de degré4 : 164.8883 + -25.77672 x + 4.364439 x^2 -0.245207 x^3 0.
004590958 x^4 ( Czech Republic Girls )
```

```
La valeur du R2 pour l'ajustement linéaire est de :
 0.8969576
La valeur du R2 pour l'ajustement de degré 2 est de :
 0.962834
a valeur du R2 pour l'ajustement de degré 3 est de :
 0.9683577
La valeur du R2 pour l'ajustement de degré 4 est de :
 0.9706672
```