Régressions

library(tidyverse)

## Error : 'format\_warning' is not an exported object from 'namespace:cli'

## -- Attaching packages -------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v ggplot2 3.2.1 v purrr 0.3.3  
## v tibble 2.1.3 v dplyr 0.8.3  
## v tidyr 1.0.0 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.4.0

## -- Conflicts ----------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

**DONNEES**

**Population totale**

setwd("~/Documents/2A/Statapp/données")  
library(readr)  
natl2015 <- read\_csv("W:/Documents/2A/Statapp/données/natl2015.csv", col\_types = cols\_only(apgar5r = col\_guess(),   
 dbwt = col\_guess(),  
 rf\_inftr = col\_guess(),  
 rf\_fedrg = col\_guess(),  
 rf\_artec = col\_guess(),  
 mager = col\_guess(),  
 mrace6 = col\_guess(),  
 dmar = col\_guess(),  
 meduc = col\_guess(),  
 fagerec11 = col\_guess(),  
 frace6 = col\_guess(),  
 feduc = col\_guess(),  
 priorlive = col\_guess(),  
 dplural = col\_guess()))  
  
head.matrix(natl2015)

## # A tibble: 6 x 14  
## mager mrace6 dmar meduc fagerec11 frace6 feduc priorlive rf\_inftr  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>   
## 1 23 41 2 2 4 4 2 2 N   
## 2 36 10 1 7 7 1 5 1 N   
## 3 19 10 1 4 2 1 4 0 N   
## 4 25 40 1 3 4 1 3 1 N   
## 5 28 10 1 6 5 1 3 0 N   
## 6 23 41 2 3 3 4 2 3 N   
## # ... with 5 more variables: rf\_fedrg <chr>, rf\_artec <chr>,  
## # apgar5r <dbl>, dplural <dbl>, dbwt <dbl>

# natl2015 compte 398873 observations

natl2015 <- subset(natl2015,(natl2015$mrace6 != 6 & natl2015$meduc != 9 & natl2015$fagerec11 != 11 & natl2015$frace6 != 9 & natl2015$frace6 != 6 & natl2015$feduc != 9 & natl2015$apgar5r != 5 & natl2015$dbwt != 9999 & natl2015$priorlive != 99))

**Population cible**

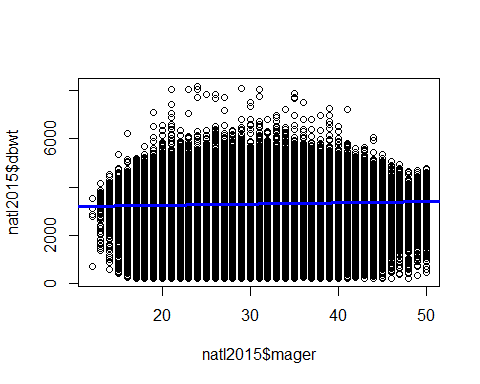
pop = subset(natl2015,rf\_inftr == "Y")  
# le jeu de données est maintenant réduit à 55922 observations

**REGRESSIONS**

**Test régression simple du poids à la naissance sur l’âge de la mère**

plot(x,y)# makea plot> abline(lm(y~ x))# plot the regressionline

LinReg=lm(natl2015$dbwt~ natl2015$mager)  
plot(natl2015$mager,natl2015$dbwt,bg="red")  
abline(LinReg,lwd=3,col="blue")

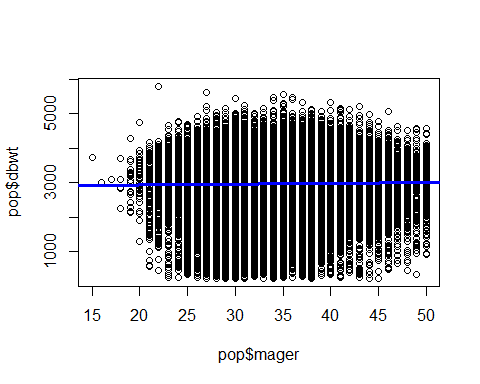


LR1 = lm(natl2015$dbwt~ natl2015$mager)  
summary(LR1)

##   
## Call:  
## lm(formula = natl2015$dbwt ~ natl2015$mager)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3144.9 -297.7 37.2 361.0 4890.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.176e+03 1.738e+00 1827.21 <2e-16 \*\*\*  
## natl2015$mager 4.132e+00 5.882e-02 70.25 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 581.4 on 3049223 degrees of freedom  
## Multiple R-squared: 0.001616, Adjusted R-squared: 0.001615   
## F-statistic: 4934 on 1 and 3049223 DF, p-value: < 2.2e-16

# poids = 4\*âge + 3175  
# coefficient significatif à 1%

LinReg2=lm(pop$dbwt~ pop$mager)  
plot(pop$mager,pop$dbwt,bg="red")  
abline(LinReg2,lwd=3,col="blue")



summary(LinReg2)

##   
## Call:  
## lm(formula = pop$dbwt ~ pop$mager)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2761.74 -443.22 99.88 531.85 2850.82   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2907.5139 21.9042 132.738 < 2e-16 \*\*\*  
## pop$mager 1.8938 0.6375 2.971 0.00297 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 773.8 on 55920 degrees of freedom  
## Multiple R-squared: 0.0001578, Adjusted R-squared: 0.0001399   
## F-statistic: 8.824 on 1 and 55920 DF, p-value: 0.002974

# y= 2\*âge + 3000  
# coefficient significatif à 1%

**Régression multiple**

MR1=lm(dbwt~ mager + meduc + mrace6 + fagerec11 + feduc + frace6 + priorlive + dmar + rf\_inftr + rf\_artec + rf\_fedrg,data=natl2015)  
summary(MR1)

##   
## Call:  
## lm(formula = dbwt ~ mager + meduc + mrace6 + fagerec11 + feduc +   
## frace6 + priorlive + dmar + rf\_inftr + rf\_artec + rf\_fedrg,   
## data = natl2015)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3193.5 -293.4 34.6 354.2 4900.9   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3400.40180 2.56649 1324.921 < 2e-16 \*\*\*  
## mager -0.67541 0.09442 -7.153 8.50e-13 \*\*\*  
## meduc 14.91874 0.28547 52.260 < 2e-16 \*\*\*  
## mrace6 -1.57941 0.04130 -38.241 < 2e-16 \*\*\*  
## fagerec11 -1.23337 0.35023 -3.522 0.000429 \*\*\*  
## feduc 10.09084 0.27022 37.344 < 2e-16 \*\*\*  
## frace6 -59.43445 0.52933 -112.283 < 2e-16 \*\*\*  
## priorlive 20.86487 0.30003 69.542 < 2e-16 \*\*\*  
## dmar -77.21698 0.81487 -94.760 < 2e-16 \*\*\*  
## rf\_inftrU -130.48034 19.68652 -6.628 3.41e-11 \*\*\*  
## rf\_inftrY -333.52142 10.14080 -32.889 < 2e-16 \*\*\*  
## rf\_artecU 40.92982 13.07721 3.130 0.001749 \*\*   
## rf\_artecX NA NA NA NA   
## rf\_artecY -81.43560 9.48751 -8.583 < 2e-16 \*\*\*  
## rf\_fedrgU NA NA NA NA   
## rf\_fedrgX NA NA NA NA   
## rf\_fedrgY 2.51769 9.30896 0.270 0.786807   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 573.6 on 3049211 degrees of freedom  
## Multiple R-squared: 0.02849, Adjusted R-squared: 0.02849   
## F-statistic: 6879 on 13 and 3049211 DF, p-value: < 2.2e-16

Tous les coefficients sont significatifs. L’âge de la mère devient correlé négativement ave le poids à la naissance, ce qui paraît cohérent. Cependant ce coefficient est relativement faible : une mère âgée d’un an de plus donne naissanceà un enfant dont le poids est 67 grammes plus faible en moyenne. (pb : l’âge n’est pas tout à fait continu, il faudrait peut-être introduire l’âge au carré, car l’âge a d’autant plus d’effet négatif que la mère est âgée) De même que l’âge du père (attention celui-ci n’est pas continu, ce qui explique le coefficient plus important). En revanche, le fait d’avoir eu un enfant avant joue positivement : si le couple (ou du moins la mère) a déjà réussi à avoir un enfant, cela signifie que le problème d’infertilité est peut-être plus facilement contournable pour elle (par exemple pas de problème au nivau génétique). Le fait de ne pas être marié joue négativement. Avoir recours à des traitements contre l’infertilité joue aussi négativement (à part pour rf\_fedrg? -> lié au fait d’avoir des jumeaux?) : en moyenne une mère ayant recours à des médicaments ou une insémination artificielle donne naissance à un nouveau-né 330 grammes plus léger.

Ajoutons l’âge de la mère au carré avant de comparer ces coefficients pour la population cible

age2 = natl2015$mager\*\*2  
MR2=lm(dbwt~ mager + age2 + meduc + mrace6 + fagerec11 + feduc + frace6 + priorlive + dmar + rf\_inftr + rf\_artec + rf\_fedrg,data=natl2015)  
summary(MR2)

##   
## Call:  
## lm(formula = dbwt ~ mager + age2 + meduc + mrace6 + fagerec11 +   
## feduc + frace6 + priorlive + dmar + rf\_inftr + rf\_artec +   
## rf\_fedrg, data = natl2015)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3193.9 -293.3 34.6 354.0 4881.5   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3011.24948 7.77420 387.339 < 2e-16 \*\*\*  
## mager 26.66217 0.52411 50.871 < 2e-16 \*\*\*  
## age2 -0.45548 0.00859 -53.027 < 2e-16 \*\*\*  
## meduc 12.90040 0.28787 44.814 < 2e-16 \*\*\*  
## mrace6 -1.53178 0.04129 -37.096 < 2e-16 \*\*\*  
## fagerec11 -2.08707 0.35044 -5.956 2.59e-09 \*\*\*  
## feduc 9.92686 0.27011 36.751 < 2e-16 \*\*\*  
## frace6 -59.78209 0.52912 -112.983 < 2e-16 \*\*\*  
## priorlive 19.51096 0.30098 64.825 < 2e-16 \*\*\*  
## dmar -70.28109 0.82493 -85.197 < 2e-16 \*\*\*  
## rf\_inftrU -130.45880 19.67745 -6.630 3.36e-11 \*\*\*  
## rf\_inftrY -328.22009 10.13662 -32.380 < 2e-16 \*\*\*  
## rf\_artecU 41.05670 13.07118 3.141 0.00168 \*\*   
## rf\_artecX NA NA NA NA   
## rf\_artecY -71.70619 9.48492 -7.560 4.03e-14 \*\*\*  
## rf\_fedrgU NA NA NA NA   
## rf\_fedrgX NA NA NA NA   
## rf\_fedrgY -0.34376 9.30483 -0.037 0.97053   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 573.3 on 3049210 degrees of freedom  
## Multiple R-squared: 0.02939, Adjusted R-squared: 0.02938   
## F-statistic: 6595 on 14 and 3049210 DF, p-value: < 2.2e-16

Avec la variable indiquant la gémelité

MR3=lm(dbwt~ mager + meduc + mrace6 + fagerec11 + feduc + frace6 + priorlive + dmar + dplural + rf\_inftr + rf\_artec + rf\_fedrg,data=natl2015)  
summary(MR3)

##   
## Call:  
## lm(formula = dbwt ~ mager + meduc + mrace6 + fagerec11 + feduc +   
## frace6 + priorlive + dmar + dplural + rf\_inftr + rf\_artec +   
## rf\_fedrg, data = natl2015)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3235.8 -291.3 23.2 336.2 4866.0   
##   
## Coefficients: (3 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4354.33404 2.94649 1477.806 < 2e-16 \*\*\*  
## mager -0.76874 0.08967 -8.573 < 2e-16 \*\*\*  
## meduc 18.22756 0.27116 67.221 < 2e-16 \*\*\*  
## mrace6 -1.62593 0.03922 -41.455 < 2e-16 \*\*\*  
## fagerec11 -0.72694 0.33260 -2.186 0.0288 \*   
## feduc 10.76124 0.25661 41.936 < 2e-16 \*\*\*  
## frace6 -60.04133 0.50268 -119.443 < 2e-16 \*\*\*  
## priorlive 31.05408 0.28547 108.781 < 2e-16 \*\*\*  
## dmar -75.43900 0.77384 -97.486 < 2e-16 \*\*\*  
## dplural -954.26919 1.65634 -576.131 < 2e-16 \*\*\*  
## rf\_inftrU -85.50169 18.69541 -4.573 4.80e-06 \*\*\*  
## rf\_inftrY -56.20665 9.64220 -5.829 5.57e-09 \*\*\*  
## rf\_artecU -0.87527 12.41895 -0.070 0.9438   
## rf\_artecX NA NA NA NA   
## rf\_artecY -8.94740 9.01067 -0.993 0.3207   
## rf\_fedrgU NA NA NA NA   
## rf\_fedrgX NA NA NA NA   
## rf\_fedrgY -20.26028 8.84032 -2.292 0.0219 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 544.7 on 3049210 degrees of freedom  
## Multiple R-squared: 0.1239, Adjusted R-squared: 0.1239   
## F-statistic: 3.079e+04 on 14 and 3049210 DF, p-value: < 2.2e-16

Maintenant, penchons nous sur les caractéristiques de la population ayant recours au traitement

MR4 = lm(dbwt~ mager + meduc + mrace6 + fagerec11 + feduc + frace6 + priorlive + dmar + rf\_artec + rf\_fedrg, data=pop)  
summary(MR4)

##   
## Call:  
## lm(formula = dbwt ~ mager + meduc + mrace6 + fagerec11 + feduc +   
## frace6 + priorlive + dmar + rf\_artec + rf\_fedrg, data = pop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2865.2 -432.3 97.5 523.2 2919.2   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2858.6810 34.3408 83.244 < 2e-16 \*\*\*  
## mager 4.8633 0.8850 5.495 3.92e-08 \*\*\*  
## meduc 12.1059 2.7986 4.326 1.52e-05 \*\*\*  
## mrace6 -0.5934 0.3968 -1.495 0.1348   
## fagerec11 1.1285 3.3969 0.332 0.7397   
## feduc 12.8916 2.4904 5.177 2.27e-07 \*\*\*  
## frace6 -56.9937 4.8775 -11.685 < 2e-16 \*\*\*  
## priorlive -48.9445 3.6620 -13.365 < 2e-16 \*\*\*  
## dmar -38.2216 18.4924 -2.067 0.0387 \*   
## rf\_artecU 42.0676 17.5360 2.399 0.0164 \*   
## rf\_artecY -96.5376 12.8171 -7.532 5.07e-14 \*\*\*  
## rf\_fedrgU NA NA NA NA   
## rf\_fedrgY 5.7968 12.4880 0.464 0.6425   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 768.3 on 55910 degrees of freedom  
## Multiple R-squared: 0.0144, Adjusted R-squared: 0.01421   
## F-statistic: 74.27 on 11 and 55910 DF, p-value: < 2.2e-16

Attention, comme on n’a gardé que ceux qui avait recours à au moins un type de traitement, la variable rf\_inftr ,e prend plus que la valeur “Y” dans ce jeu de données, il faut donc l’exclure de la régression. L’âge du père n’est pas significatif. L’âge de la mère joue positivement ? L’éducation joue positivement à des niveaux équivalents pour le père et la mère, contrairement à précédemment. Ne pas être marié joue négativment mais moins. Le coefficient de priorlive est étonant. Encore une fois

age3 = pop$mager\*\*2  
MR5 = lm(dbwt~ mager + meduc + mrace6 + age3 + fagerec11 + feduc + frace6 + priorlive + dmar + rf\_artec + rf\_fedrg, data=pop)  
summary(MR5)

##   
## Call:  
## lm(formula = dbwt ~ mager + meduc + mrace6 + age3 + fagerec11 +   
## feduc + frace6 + priorlive + dmar + rf\_artec + rf\_fedrg,   
## data = pop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2874.47 -433.15 96.91 522.66 2931.53   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2355.68057 110.48391 21.321 < 2e-16 \*\*\*  
## mager 34.80865 6.31414 5.513 3.55e-08 \*\*\*  
## meduc 10.24245 2.82496 3.626 0.000288 \*\*\*  
## mrace6 -0.56815 0.39676 -1.432 0.152158   
## age3 -0.42662 0.08907 -4.790 1.67e-06 \*\*\*  
## fagerec11 0.56142 3.39827 0.165 0.868781   
## feduc 12.41559 2.49190 4.982 6.30e-07 \*\*\*  
## frace6 -57.53704 4.87788 -11.796 < 2e-16 \*\*\*  
## priorlive -49.49021 3.66309 -13.511 < 2e-16 \*\*\*  
## dmar -30.59986 18.55709 -1.649 0.099162 .   
## rf\_artecU 41.51305 17.53292 2.368 0.017901 \*   
## rf\_artecY -97.88576 12.81765 -7.637 2.26e-14 \*\*\*  
## rf\_fedrgU NA NA NA NA   
## rf\_fedrgY 5.42243 12.48575 0.434 0.664080   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 768.2 on 55909 degrees of freedom  
## Multiple R-squared: 0.01481, Adjusted R-squared: 0.01459   
## F-statistic: 70.02 on 12 and 55909 DF, p-value: < 2.2e-16

MR6 = lm(dbwt~ mager + meduc + mrace6 + age3 + fagerec11 + feduc + frace6 + priorlive + dmar + dplural + rf\_artec + rf\_fedrg, data=pop)  
summary(MR6)

##   
## Call:  
## lm(formula = dbwt ~ mager + meduc + mrace6 + age3 + fagerec11 +   
## feduc + frace6 + priorlive + dmar + dplural + rf\_artec +   
## rf\_fedrg, data = pop)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3212.5 -298.1 55.9 385.8 3109.9   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3537.11631 87.98681 40.201 < 2e-16 \*\*\*  
## mager 37.45772 5.01459 7.470 8.15e-14 \*\*\*  
## meduc 13.47388 2.24360 6.005 1.92e-09 \*\*\*  
## mrace6 -0.97439 0.31511 -3.092 0.001987 \*\*   
## age3 -0.56215 0.07074 -7.947 1.95e-15 \*\*\*  
## fagerec11 4.86094 2.69894 1.801 0.071700 .   
## feduc 10.63872 1.97905 5.376 7.66e-08 \*\*\*  
## frace6 -56.15226 3.87392 -14.495 < 2e-16 \*\*\*  
## priorlive 58.78861 2.97008 19.794 < 2e-16 \*\*\*  
## dmar -54.70704 14.73828 -3.712 0.000206 \*\*\*  
## dplural -903.16341 4.99185 -180.927 < 2e-16 \*\*\*  
## rf\_artecU 3.29592 13.92590 0.237 0.812909   
## rf\_artecY -13.85930 10.19012 -1.360 0.173812   
## rf\_fedrgU NA NA NA NA   
## rf\_fedrgY -16.61997 9.91669 -1.676 0.093752 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 610.1 on 55908 degrees of freedom  
## Multiple R-squared: 0.3786, Adjusted R-squared: 0.3785   
## F-statistic: 2621 on 13 and 55908 DF, p-value: < 2.2e-16

**Annexe** Corélation entre le recours à médicaments et la gémélité

gemelite = lm(dplural~ mager + rf\_fedrg + rf\_artec, data=natl2015)  
summary(gemelite)

##   
## Call:  
## lm(formula = dplural ~ mager + rf\_fedrg + rf\_artec, data = natl2015)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.4154 -0.0353 -0.0306 -0.0247 3.9671   
##   
## Coefficients: (2 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.282e+00 3.393e-03 377.871 < 2e-16 \*\*\*  
## mager 1.184e-03 1.924e-05 61.506 < 2e-16 \*\*\*  
## rf\_fedrgU -1.012e-01 4.092e-03 -24.730 < 2e-16 \*\*\*  
## rf\_fedrgX -2.859e-01 3.336e-03 -85.697 < 2e-16 \*\*\*  
## rf\_fedrgY -2.363e-02 3.063e-03 -7.714 1.22e-14 \*\*\*  
## rf\_artecU NA NA NA NA   
## rf\_artecX NA NA NA NA   
## rf\_artecY 7.410e-02 3.122e-03 23.733 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1888 on 3049219 degrees of freedom  
## Multiple R-squared: 0.05037, Adjusted R-squared: 0.05036   
## F-statistic: 3.234e+04 on 5 and 3049219 DF, p-value: < 2.2e-16