Data Preprocessing

```
births <- read.csv("chds_births.csv")
meth.names <- c('Caucasian','Caucasian','Caucasian','Caucasian','Caucasian','Caucasian', 'Mexican', 'Af
med.names <- c('elementary', 'middle', 'hs', 'hs + trade', 'hs + college', 'college', 'trade', 'unclear
feth.names <- c('Caucasian','Caucasian','Caucasian','Caucasian','Caucasian','Caucasian', 'Mexican', 'Af
fed.names <- c('elementary', 'middle', 'hs', 'hs + trade', 'hs + college', 'college', 'trade', 'unclear
marital.names <- c(NA, 'married', 'separated', 'divorced', 'widowed', 'never married')
income.names <- c('<2500', '2500-4999', '5000-7499', '7500-9999', '10000-12499', '12500-14999', '15000-
smoke.names <- c('never', 'now', 'until pregnancy', 'used to')
time.names <- c('never', 'still smokes', 'during pregnancy', 'less than a year', '1-2yrs', '2-3yrs', '3-
number.names <- c('never', '1-4', '5-9', '10-14', '15-19', '20-29', '30-39', '40-60', '>60', 'smoked, a

births$meth <- meth.names[births$meth + 1]
births$feth<- feth.names[births$feth + 1]
births$fed <- fed.names[births$fed + 1]
births$marital <- marital.names[births$marital+1]
births$income <- income.names[births$income + 1]
births$smoke <- smoke.names[births$smoke + 1]
births$time <- time.names[births$time + 1]
births$number <- number.names[births$number + 1]
summary(births)
```

```
##        wt          gestation        parity           meth
##  Min.   : 55.0   Min.   :148.0   Min.   : 0.000   Length:1236
##  1st Qu.:108.8   1st Qu.:272.0   1st Qu.: 0.000   Class :character
##  Median :120.0   Median :280.0   Median : 1.000   Mode  :character
##  Mean   :119.6   Mean   :279.3   Mean   : 1.932
##  3rd Qu.:131.0   3rd Qu.:288.0   3rd Qu.: 3.000
##  Max.   :176.0   Max.   :353.0   Max.   :13.000
##                  NA's   :13
##       mage            med              mht             mwt
##  Min.   :15.00   Min.   :0.000   Min.   :53.00   Min.   : 87.0
##  1st Qu.:23.00   1st Qu.:2.000   1st Qu.:62.00   1st Qu.:114.8
##  Median :26.00   Median :2.000   Median :64.00   Median :125.0
##  Mean   :27.26   Mean   :2.917   Mean   :64.05   Mean   :128.6
##  3rd Qu.:31.00   3rd Qu.:4.000   3rd Qu.:66.00   3rd Qu.:139.0
##  Max.   :45.00   Max.   :7.000   Max.   :72.00   Max.   :250.0
##  NA's   :2       NA's   :1       NA's   :22      NA's   :36
##      feth                fage           fed                  fht
##  Length:1236        Min.   :18.00   Length:1236        Min.   :60.0
##  Class :character   1st Qu.:25.00   Class :character   1st Qu.:68.0
##  Mode  :character   Median :29.00   Mode  :character   Median :71.0
##                     Mean   :30.35                      Mean   :70.2
##                     3rd Qu.:34.00                      3rd Qu.:72.0
##                     Max.   :62.00                      Max.   :78.0
##                     NA's   :7                          NA's   :492
##       fwt          marital               income              smoke
##  Min.   :110.0   Length:1236        Length:1236        Length:1236
##  1st Qu.:155.0   Class :character   Class :character   Class :character
##  Median :170.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   :171.2
##  3rd Qu.:185.0
##  Max.   :260.0
##  NA's   :499
##      time              number
##  Length:1236        Length:1236
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Initial Model

Variables to definitely include

1. gestation
2. parity
3. time
4. number
5. smoke
6. martial
7. fed
8. med

Variables to consider including

1. meth
2. feth
3. mage/fage (not both – correlated)
4. mht/mwt (slightly correlated so probably not both – mwt might be better)

Non-linear effects/ other modifications to covariates

- change grouping of smoke: group 0 and 3 together; 1 and 2 together to form "never/used to" and "now/until pregnancy"
- change grouping of med/fed: (0, 1, 7) becomes group "no highschool/ highschool unclear", (3, 6) -> trade, (4, 5) -> college [the latter 2 groupings are relevant for fed more than med]
- income* (have to fix with imputation first)

Forward Selection using just the variables in “to include”

```
keeps <- c("wt", "gestation", "parity", "time", "number", "smoke", "marital", "fed", "med", "mwt", "mht
birth.data <- births[keeps]
birth.data <- na.omit(birth.data)
print(dim(birth.data))
```

```
## [1] 1152   11
```

```
M0 <- lm(wt ~ 1, data = birth.data)
Mmax <- lm(wt ~ gestation + parity + time + number+ smoke + fed, data=birth.data)
Mstart <- lm(wt ~ gestation + parity + smoke + fed, data=birth.data)
MgestPar <- lm(wt ~ gestation + parity + I(gestation*parity) + smoke + fed, data = birth.data)
Mbio1 <- lm(wt ~ gestation + parity + mwt + mht + I(mwt*mht^2) + I(gestation*parity) + time, data = bir
Mbio2 <- lm(wt ~ gestation + parity + mwt + mht + I(mwt*mht^2) + I(gestation*parity) + smoke, data = bi
Mbio3 <- lm(wt ~ gestation + parity + mwt + mht + I(mwt*mht^2) + I(gestation*parity) + number, data = b
MbioBMI <- lm(wt ~ gestation + parity + mwt + mht + I(mwt*703/mht^2) + I(gestation*parity) + number, da
bmi <- birth.data$mwt*703/(birth.data$mht)^2
bmiLB <- bmi - 18.5
bmiUB <- 24.9 - bmi
birth.data$bmiLB <- bmiLB
birth.data$bmiUB <- bmiUB
MbioBMICenteredUB <- lm(wt ~ gestation + parity + mwt + mht + bmiUB + number, data = birth.data)
MbioBMICenteredLB <- lm(wt ~ gestation + parity + mwt + mht + bmiLB + number, data = birth.data)
MbioBMICentered <- lm(wt ~ gestation + parity + I(gestation*parity) + bmiUB + number, data = birth.data
ntot <- dim(birth.data)[1]
ntrain <- 1000
set.seed(5)
train.ind <- sample(ntot, ntrain)

M0 <- update(M0, subset = train.ind)
Mmax <- update(Mmax, subset = train.ind)
Mstart <- update(Mstart, subset = train.ind)
```

```
# forward selection
Mfwd <- step(object = M0, # starting point model
scope = list(lower = M0, upper = Mmax), # smallest and largest model
direction = "forward",
trace = FALSE) # trace prints out information
print(Mfwd$call)
```

```
## lm(formula = wt ~ gestation + smoke + parity, data = birth.data,
##     subset = train.ind)
```

```
# backward elimiation
Mback <- step(object = Mmax, # starting point model
scope = list(lower = M0, upper = Mmax),
direction = "backward", trace = FALSE)
print(Mback$call)
```

```
## lm(formula = wt ~ gestation + parity + time, data = birth.data,
##     subset = train.ind)
```

```
# stepwise selection (both directions)
Mstep <- step(object = Mstart,
scope = list(lower = M0, upper = Mmax),
direction = "both", trace = FALSE)
print(Mstep$call)
```

```
## lm(formula = wt ~ gestation + parity + smoke, data = birth.data,
##     subset = train.ind)
```

The MSPE for training set of 1000, seed=5, omit NA is

```
print(Mfwd$call)
```

```
## lm(formula = wt ~ gestation + smoke + parity, data = birth.data,
##     subset = train.ind)
```

```
print(sum((birth.data$wt[-train.ind] - predict(Mfwd, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 46351.24
```

```
print(Mback$call)
```

```
## lm(formula = wt ~ gestation + parity + time, data = birth.data,
##     subset = train.ind)
```

```
print(sum((birth.data$wt[-train.ind] - predict(Mback, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 45870.46
```

```
print(Mstep$call)
```

```
## lm(formula = wt ~ gestation + parity + smoke, data = birth.data,
##     subset = train.ind)
```

```
print(sum((birth.data$wt[-train.ind] - predict(Mstep, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 46351.24
```

```
print(Mstart$call)
```

```
## lm(formula = wt ~ gestation + parity + smoke + fed, data = birth.data,
##     subset = train.ind)
```

```
print(sum((birth.data$wt[-train.ind] - predict(Mstart, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 46739.36
```

```
print(MgestPar$call)
```

```
## lm(formula = wt ~ gestation + parity + I(gestation * parity) +
##     smoke + fed, data = birth.data)
```

```
print(sum((birth.data$wt[-train.ind] - predict(MgestPar, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 46193.63
```

```
print(Mbio1$call)
```

```
## lm(formula = wt ~ gestation + parity + mwt + mht + I(mwt * mht^2) +
##     I(gestation * parity) + time, data = birth.data)
```

```
print(sum((birth.data$wt[-train.ind] - predict(Mbio1, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 43103.81
```

```
print(Mbio2$call)
```

```
## lm(formula = wt ~ gestation + parity + mwt + mht + I(mwt * mht^2) +
##     I(gestation * parity) + smoke, data = birth.data)
```

```r
print(sum((birth.data$wt[-train.ind] - predict(Mbio2, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 43786.71
```

```r
print(Mbio3$call)
```

```
## lm(formula = wt ~ gestation + parity + mwt + mht + I(mwt * mht^2) +
##     I(gestation * parity) + number, data = birth.data)
```

```r
print(sum((birth.data$wt[-train.ind] - predict(Mbio3, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 42554.74
```

```r
print(MbioBMI$call)
```

```
## lm(formula = wt ~ gestation + parity + mwt + mht + I(mwt * 703/mht^2) +
##     I(gestation * parity) + number, data = birth.data)
```

```r
print(sum((birth.data$wt[-train.ind] - predict(MbioBMI, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 42058.56
```

```r
print(MbioBMICentered$call)
```

```
## lm(formula = wt ~ gestation + parity + I(gestation * parity) +
##     bmiUB + number, data = birth.data)
```

```r
print(sum((birth.data$wt[-train.ind] - predict(MbioBMICentered, newdata = birth.data[-train.ind,]))^2))
```

```
## [1] 45739.36
```

```r
print(MbioBMICenteredLB$call)
```

```
## lm(formula = wt ~ gestation + parity + mwt + mht + bmiLB + number,
##     data = birth.data)
```

```r
print(sum((birth.data$wt[-train.ind] - predict(MbioBMICenteredLB, newdata = birth.data[-train.ind,]))^2
```

```
## [1] 41928.23
```

```r
print(MbioBMICenteredUB$call)
```

```
## lm(formula = wt ~ gestation + parity + mwt + mht + bmiUB + number,
##     data = birth.data)
```

```r
print(sum((birth.data$wt[-train.ind] - predict(MbioBMICenteredUB, newdata = birth.data[-train.ind,]))^2
```

```
## [1] 41928.23
```