Group Analysis

Alex Schulte, Christine Lo, Matthew Reyes, Alexander Adia

2023-04-06

Research Question

Existing studies find that infants born at low birth weight (LBW) are at an increased risk of physical disabilities and impaired cognitive development. While genetic factors contribute to LBW, maternal smoking during pregnancy has been identified as the most significant modifiable risk factor. We seek to answer the following question: what is the effect of maternal smoking during pregnancy on the likelihood of having a LBW infant?

The target population for this study is live singleton first births in the US in 2015. We are limiting the population to singleton first births because multiples are associated with lower birth weight, and infants from subsequent pregnancies have been shown to have higher birth weights than those from first pregnancies.

Target Causal Parameter

We aim to estimate the causal risk difference: $\Psi^*(P^*) = P^*(Y1-1) - P^*(Y0-1) = E^*(Y1) - E^*(Y0)$

The target causal parameter is the difference in the counterfactual risk of LBW if all expectant mothers in the population smoked during pregnancy vs. if all expectant mothers in the population did not smoke during pregnancy.

Data Set

First, we import the data set for 2015 and inspect it, including variables available.

```
## [1] "D:/MPH/causal_252_group"
```

```
##   [1] "laterec"     "dob_yy"      "dob_mm"      "dob_tt"      "dob_wk"
##   [6] "bfacil"      "f_bfacil"    "bfacil3"     "mageimp"     "magerep"
##  [11] "mager"       "mager14"     "mager9"      "mbstate_rec" "restatus"
##  [16] "mrace31"     "mrace6"      "mrace15"     "mbrace"      "mraceimp"
##  [21] "mhisp_r"     "f_mhisp"     "mracehisp"   "mar_p"       "dmar"
##  [26] "mar_imp"     "f_mar_p"     "meduc"       "f_meduc"     "fagerpt_flg"
##  [31] "fagecomb"    "fage11"      "frace31"     "frace6"      "frace15"
##  [36] "fbrace"      "fhisp_r"     "f_fhisp"     "fracehisp"   "feduc"
##  [41] "f_feduc"     "riorlive"    "riordead"    "riorterm"    "lbo_rec"
##  [46] "tbo_rec"     "illb_r"      "illb_r11"    "ilop_r"      "ilop_r11"
##  [51] "ilp_r"       "ilp_r11"     "recare"      "f_mpcb"      "recare5"
##  [56] "revis"       "revis_rec"   "f_tpcv"      "wic"         "f_wic"
##  [61] "cig_0"       "cig_1"       "cig_2"       "cig_3"       "cig0_r"
##  [66] "cig1_r"      "cig2_r"      "cig3_r"      "f_cigs_0"    "f_cigs_1"
```

```
##  [71] "f_cigs_2"    "f_cigs_3"    "cig_rec"     "f_tobaco"    "mhtr"
##  [76] "f_m_ht"      "bmi"         "bmi_r"       "wgt_r"       "f_pwgt"
##  [81] "dwgt_r"      "f_dwgt"      "wtgain"      "wtgain_rec"  "f_wtgain"
##  [86] "rf_pdiab"    "rf_gdiab"    "rf_phype"    "rf_ghype"    "rf_ehype"
##  [91] "rf_ppb"      "f_rf_pdiab"  "f_rf_gdiab"  "f_rf_phype"  "f_rf_ghype"
##  [96] "f_rf_ehype"  "f_rf_ppb"    "rf_inft"     "rf_drg"      "rf_art"
## [101] "f_rf_drg"    "f_rf_art"    "rf_cesar"    "rf_cesarn"   "f_rf_cesar"
## [106] "f_rf_ncesar" "no_risks"    "ip_gon"      "ip_syph"     "ip_chlam"
## [111] "ip_hepb"     "ip_hepc"     "f_ip_gon"    "f_ip_syph"   "f_ip_chlam"
## [116] "f_ip_hepb"   "f_ip_hepc"   "no_infec"    "ob_succ"     "ob_fail"
## [121] "f_ob_succ"   "f_ob_fail"   "seqnum_co"   "ld_indl"     "ld_augm"
## [126] "ld_ster"     "ld_antb"     "ld_chor"     "ld_anes"     "f_ld_indl"
## [131] "f_ld_augm"   "f_ld_ster"   "f_ld_antb"   "f_ld_chor"   "f_ld_anes"
## [136] "no_lbrdlv"   "me_pres"     "me_rout"     "me_trial"    "f_me_pres"
## [141] "f_me_rout"   "f_me_trial"  "rdmeth_rec"  "dmeth_rec"   "f_dmeth_rec"
## [146] "mm_mtr"      "mm_plac"     "mm_rupt"     "mm_uhyst"    "mm_aicu"
## [151] "f_mm_mtr"    "f_mm_"       "f_mm_rupt"   "f_mm_uhyst"  "f_mm_aicu"
## [156] "no_mmorb"    "attend"      "mtran"       "ay"          "ay_rec"
## [161] "f_pay"       "f_pay_rec"   "apgar5"      "apgar5r"     "f_apgar5"
## [166] "apgar10"     "apgar10r"    "dplural"     "imp_plur"    "setorder_r"
## [171] "sex"         "imp_sex"     "dlmp_mm"     "dlmp_yy"     "combgst_imp"
## [176] "obgest_flg"  "combgest"    "estrec10"    "estrec3"     "lmpused"
## [181] "oegest_comb" "oegest_r10"  "oegest_r3"   "bwtr14"      "bwtr4"
## [186] "brthwgt"     "bwtimp"      "ab_aven1"    "ab_aven6"    "ab_nicu"
## [191] "ab_surf"     "ab_anti"     "ab_seiz"     "f_ab_aven1"  "f_ab_aven6"
## [196] "f_ab_nicu"   "f_ab_surf"   "f_ab_anti"   "f_ab_seiz"   "no_abnorm"
## [201] "ca_anen"     "ca_mnsb"     "ca_cchd"     "ca_cdh"      "ca_omph"
## [206] "ca_gast"     "f_ca_anen"   "f_ca_mnsb"   "f_ca_cchd"   "f_ca_cdh"
## [211] "f_ca_omph"   "f_ca_gast"   "ca_limb"     "ca_cleft"    "ca_clpal"
## [216] "ca_down"     "ca_disor"    "ca_hypo"     "f_ca_limb"   "f_ca_cleft"
## [221] "f_ca_clpal"  "f_ca_down"   "f_ca_disor"  "f_ca_hypo"   "no_congen"
## [226] "itran"       "ilive"       "bfed"        "f_bfed"      "ubfacil"
## [231] "urf_diab"    "urf_chype"   "urf_phype"   "urf_ehype"   "ume_forc"
## [236] "ume_vacu"    "uob_indu"    "uld_bree"    "uca_anen"    "uca_spina"
## [241] "uca_omph"    "uca_clip"    "uca_hern"    "uca_down"    "flgnd"
## [246] "aged"        "ager5"       "ager22"      "manner"      "dispo"
## [251] "autopsy"     "lace"        "ucod"        "ucodr130"    "recwt"
```

Next, we select variables of interest for our analysis by subsetting from the larger data set.

Variables included in the dataset are as follows:

Variable Name Type Descriptive summary of measure

smoked Exposure (A, binary) This variable is considered the intervention or exposure of interest - it's a measure of whether the mother was considered a smoker (at least 1 cigarette/day) during any of the three trimesters.

lbw Outcome (Y, binary) This variable is the outcome, which is the weight of the infant at time of birth, classified as low birth weight (1) when the birthweight was below 2500 grams. Birth weight greater than 2500 grams is coded as 0.

mrace15 Endogenous covariate Categorical race variable

mhisp_r Endogenous covariate Categorical variable indicating hispanic origin status

mager9 Endogenous covariate Bins of age ranges

dmar Endogenous covariate Categorical variable of marital status

meduc Endogenous covariate Categorical variable of mother's achieved education level

wic Endogenous covariate Indicator variable of mother receiving WIC benefits

mhtr Endogenous covariate Continuous variable of mother's height

bmi Endogenous covariate Continuous variable of mother's bmi

dwgt_r Endogenous covariate Continuous variable of mother's weight at time of birth

rf_pdiab Endogenous covariate Indicator variable of mother having pre-pregnancy diabetes

rf_gdiab Endogenous covariate Indicator variable of mother with gestational diabetes

rf_phype Endogenous covariate Indicator variable of mother with pre-pregnancy hypertension

ip_gon Endogenous covariate Indicator variable of gonorrhea infection at time of birth

ip_syph Endogenous covariate Indicator variable of syphilis infection at time of birth

ip_chlam Endogenous covariate Indicator variable of chlamydia infection at time of birth

ip_hepb Endogenous covariate Indicator variable of Hepatitis B infection at time of birth

ip_hepc Endogenous covariate Indicator variable of Hepatitis C infection at time of birth

oegest_comb Endogenous covariate An edited obstetric estimate of weeks of gestation, discrete 17-47

tbo_rec Endogenous covariate Continuous variable of birth order

wtgain Endogenous covariate Continuous variable of mother's weight gain during gestation

precare5 Endogenous covariate Categorical variable of when prenatal care began

sex Endogenous covariate Categorical variable of sex of infant

Data Cleaning

Then, we recode some of the variables of interest into outcome and exposure variables A and Y. We also prepare the covariates and endogenous variables for analysis by recoding them into indicator or dummy variables. We also remove missings or unknowns, which is a very conservative analysis approach - future analysis may utilize data imputation, but given the large number of records in this data set and the relatively small number of missing/unknown data, for the purpose of this assignment the more conservative approach is taken.

```
## # A tibble: 6 x 26
##   sex   tbo_rec illb_r ilop_r ilp_r mrace15 mhisp_r mager9  dmar meduc   wic
##   <chr>   <dbl>  <dbl>  <dbl> <dbl>   <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 F           3     12    888    12      10       0      3     2     2     1
## 2 F           2     92    888    92       6       0      4     1     3     0
## 3 F           1    888    888   888       1       0      4     1     6     0
## 4 F           3     58    888    58       1       0      3     1     2     1
## 5 F           1    888    888   888       1       0      5     1     6     0
## 6 F           1    888    888   888      15       0      4     1     6     0
## # ... with 15 more variables: mhtr <dbl>, bmi <dbl>, dwgt_r <dbl>,
## #   wtgain <dbl>, rf_pdiab <dbl>, rf_gdiab <dbl>, rf_phype <dbl>, ip_gon <dbl>,
## #   ip_syph <dbl>, ip_chlam <dbl>, ip_hepb <dbl>, ip_hepc <dbl>,
## #   oegest_comb <dbl>, lbw <dbl>, smoked <dbl>
```

The dataset consists of a number of variables describing births and pregnancies in the United States in the year 2015 with 3245637 records after removing missing values, obtained from the National Center for Health Statistics.

Marginal Distributions of Exposure and Outcome

For the mothers smoking status during the pregnancy, we observe: c(0, 1), c(3009817, 235820)

For the low birth weight status at time of birth (outcome), we observe: c(0, 1), c(2996200, 249437)

Expected Challenges

Anticipated challenges include:

Identifying singleton births - we may need to create a unique identifier for each mother Computational strain given the size of the data Remaining potential for uncontrolled confounding (e.g., genetics, traumatic experiences during pregnancy)

Expected Deviations

Potentially measuring only singleton births (single live birth per delivery) or even first live births from that mother.

Analysis

G-Comp

IPW

SuperLearner/TMLE

Works Cited

Almond, Douglas, Kenneth Y. Chay and David S. Lee. "The Costs Of Low Birth Weight," Quarterly Journal of Economics, 2005, v120(3,Aug), 1031-1083.

Bacci S, Bartolucci F, Chiavarini M, Minelli L, Pieroni L. Differences in birthweight outcomes: a longitudinal study based on siblings. Int J Environ Res Public Health. 2014 Jun;11(6):6472-84. doi: 10.3390/ijerph110606472. PMID: 25003169; PMCID: PMC4076673.

Bohn C, Vogel M, Poulain T et al. Birth weight increases with birth order despite decreasing maternal pregnancy weight gain. Acta Paediatr 2021;110:1218–24.

National Center for Health Statistics (2015). Data File Documentations, Birth Cohort Linked Birth/Infant Death, 2015, National Center for Health Statistics, Hyattsville, Maryland. https://www.nber.org/research/data/linked-birthinfant-death-cohort-data

Marginal Distribution Tables

```
## # A tibble: 2 x 2
##   sex         n
##   <chr>   <int>
```

```
## 1 F     1583823
## 2 M     1661814
## # A tibble: 8 x 2
##   tbo_rec       n
##     <dbl>   <int>
## 1       1 1161621
## 2       2  938887
## 3       3  567722
## 4       4  294621
## 5       5  142428
## 6       6   67738
## 7       7   33833
## 8       8   38787
## # A tibble: 301 x 2
##    illb_r     n
##     <dbl> <int>
##  1      0 62766
##  2      1  3063
##  3      3 42479
##  4      4   341
##  5      5   391
##  6      6   619
##  7      7   777
##  8      8   938
##  9      9  2279
## 10     10  6091
## # ... with 291 more rows
## # A tibble: 301 x 2
##    ilop_r     n
##     <dbl> <int>
##  1      0 62766
##  2      1  3063
##  3      3   467
##  4      4   269
##  5      5   377
##  6      6   614
##  7      7  1090
##  8      8  2383
##  9      9 12579
## 10     10 25386
## # ... with 291 more rows
## # A tibble: 301 x 2
##    ilp_r     n
##    <dbl> <int>
##  1     0 62766
##  2     1  3063
##  3     3 42866
##  4     4   606
##  5     5   748
##  6     6  1199
##  7     7  1817
##  8     8  3189
##  9     9 14483
## 10    10 30944
```

```
## # ... with 291 more rows
## # A tibble: 15 x 2
##    mrace15       n
##      <dbl>   <int>
##  1       1 2470333
##  2       2  458194
##  3       3   31019
##  4       4   57108
##  5       5   43921
##  6       6   27441
##  7       7    5786
##  8       8   12734
##  9       9   17362
## 10      10   36231
## 11      11     862
## 12      12    1137
## 13      13    1754
## 14      14    7657
## 15      15   74098
## # A tibble: 6 x 2
##   mhisp_r       n
##     <dbl>   <int>
## 1       0 2475268
## 2       1  475329
## 3       2   53072
## 4       3   15908
## 5       4  109817
## 6       5  116243
## # A tibble: 9 x 2
##   mager9      n
##    <dbl>  <int>
## 1      1   2243
## 2      2 204620
## 3      3 719710
## 4      4 947648
## 5      5 879722
## 6      6 404645
## 7      7  81048
## 8      8   5552
## 9      9    449
## # A tibble: 2 x 2
##    dmar       n
##   <dbl>   <int>
## 1     1 1963572
## 2     2 1282065
## # A tibble: 9 x 2
##   meduc      n
##   <dbl>  <int>
## 1     0  62766
## 2     1 109758
## 3     2 343629
## 4     3 790383
## 5     4 673892
## 6     5 261129
```

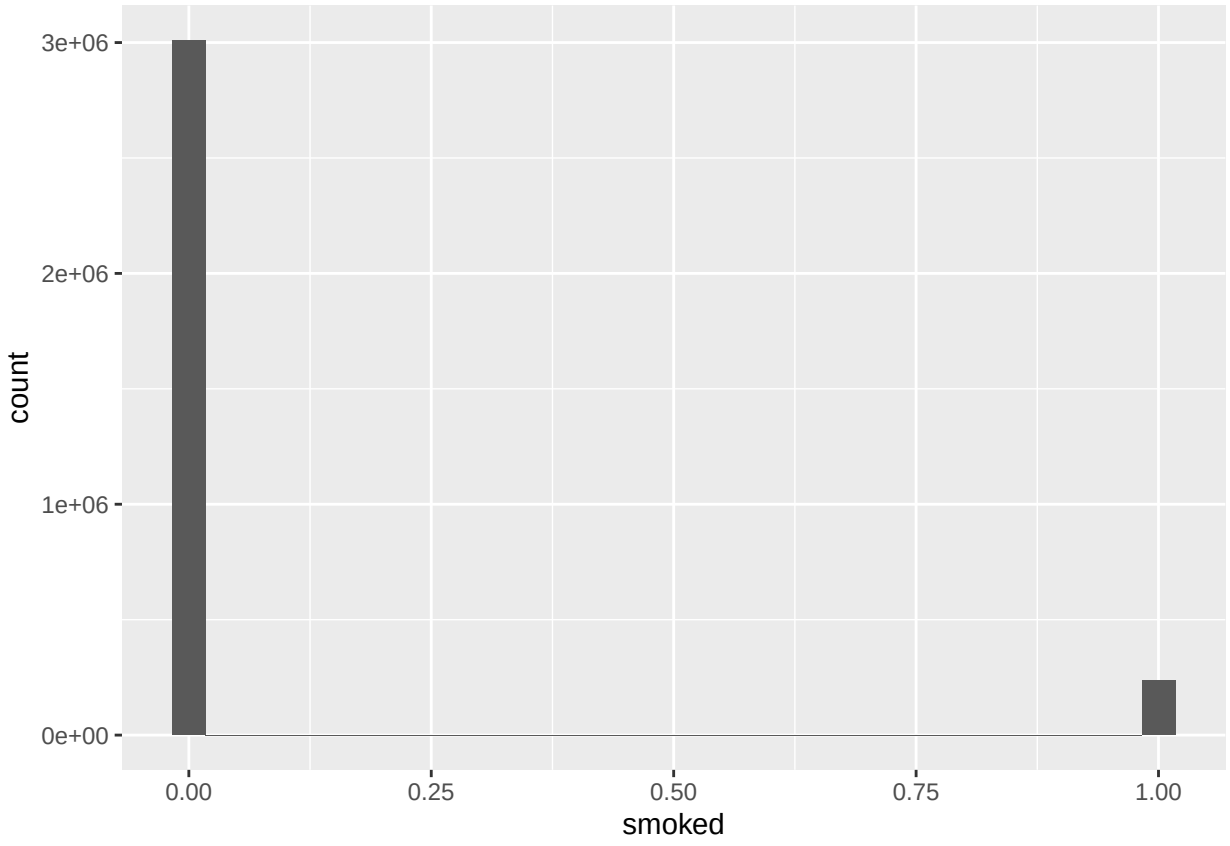
```
## 7     6 637766
## 8     7 284378
## 9     8  81936
## # A tibble: 2 x 2
##     wic       n
##   <dbl>   <int>
## 1     0 1935078
## 2     1 1310559
## # A tibble: 47 x 2
##     mhtr     n
##    <dbl> <int>
##  1     0 62766
##  2     1  3063
##  3    30    17
##  4    35     1
##  5    36     8
##  6    37     1
##  7    38    10
##  8    39     6
##  9    40    10
## 10    41     4
## # ... with 37 more rows
## # A tibble: 562 x 2
##      bmi     n
##    <dbl> <int>
##  1   0   62766
##  2   1    3063
##  3  13      12
##  4  13.1    21
##  5  13.2    30
##  6  13.3    45
##  7  13.4    40
##  8  13.5    35
##  9  13.6    57
## 10  13.7    91
## # ... with 552 more rows
## # A tibble: 400 x 2
##    dwgt_r     n
##     <dbl> <int>
##  1      0   779
##  2      1   108
##  3      2   177
##  4      3   179
##  5      4   205
##  6      5   296
##  7      6   263
##  8      7   283
##  9      8   341
## 10      9   331
## # ... with 390 more rows
## # A tibble: 99 x 2
##    wtgain     n
##     <dbl> <int>
##  1      0 84138
```

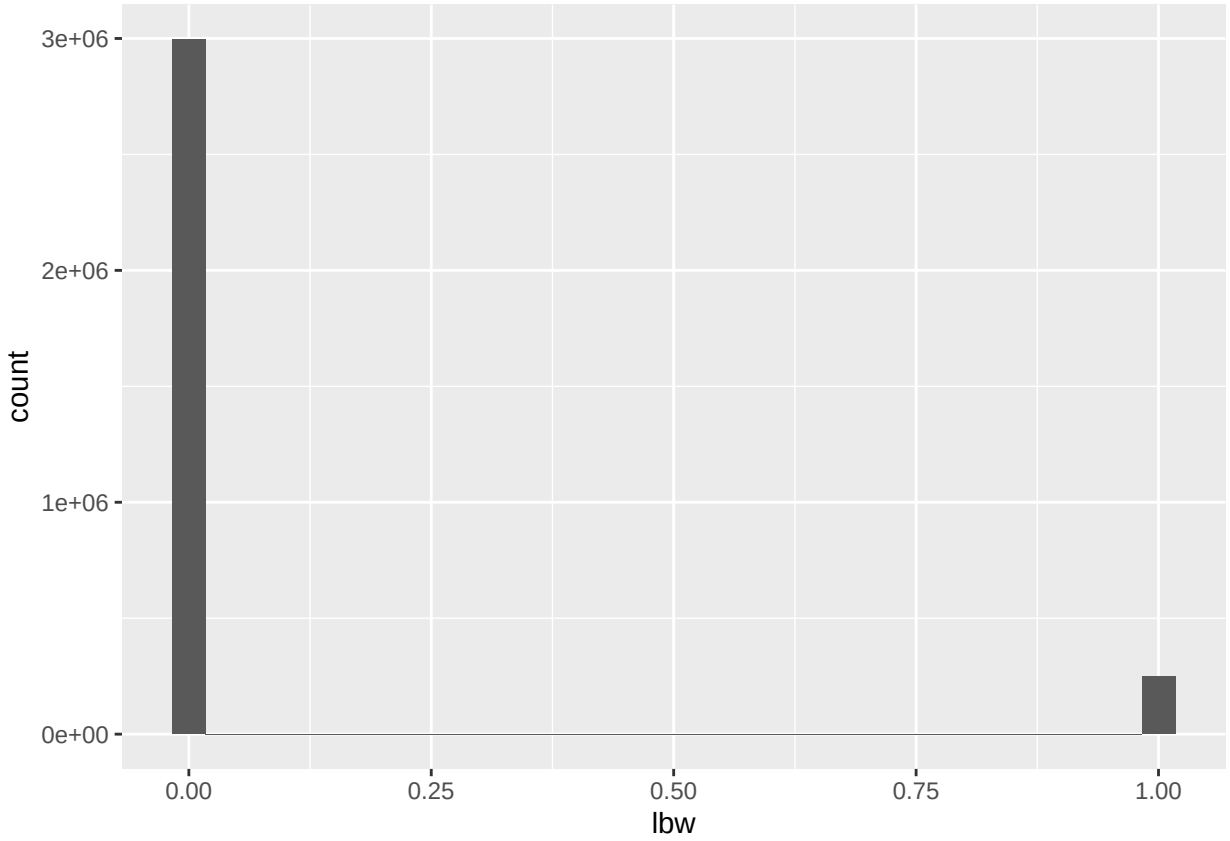
```
##  2      1 10112
##  3      2 12755
##  4      3 13323
##  5      4 14829
##  6      5 20171
##  7      6 19032
##  8      7 20489
##  9      8 23360
## 10      9 23250
## # ... with 89 more rows
## # A tibble: 2 x 2
##   rf_pdiab       n
##      <dbl>   <int>
## 1        0 3221273
## 2        1   24364
## # A tibble: 2 x 2
##   rf_gdiab       n
##      <dbl>   <int>
## 1        0 3068577
## 2        1  177060
## # A tibble: 2 x 2
##   rf_phype       n
##      <dbl>   <int>
## 1        0 3196979
## 2        1   48658
## # A tibble: 2 x 2
##   ip_gon       n
##    <dbl>   <int>
## 1      0 3237557
## 2      1    8080
## # A tibble: 2 x 2
##   ip_syph       n
##     <dbl>   <int>
## 1       0 3243300
## 2       1    2337
## # A tibble: 2 x 2
##   ip_chlam       n
##      <dbl>   <int>
## 1        0 3186976
## 2        1   58661
## # A tibble: 2 x 2
##   ip_hepb       n
##     <dbl>   <int>
## 1       0 3239159
## 2       1    6478
## # A tibble: 2 x 2
##   ip_hepc       n
##     <dbl>   <int>
## 1       0 3235572
## 2       1   10065
## # A tibble: 31 x 2
##    oegest_comb     n
##          <dbl> <int>
##  1          17   147
```

```
##  2          18   244
##  3          19   424
##  4          20   718
##  5          21  1040
##  6          22  1297
##  7          23  1943
##  8          24  2687
##  9          25  2978
## 10          26  3395
## # ... with 21 more rows
## # A tibble: 2 x 2
##     lbw       n
##   <dbl>   <int>
## 1     0 2996200
## 2     1  249437
## # A tibble: 2 x 2
##   smoked       n
##    <dbl>   <int>
## 1      0 3009817
## 2      1  235820
```

```
##      sex                tbo_rec          illb_r          ilop_r
##  Length:3245637     Min.   :1.000   Min.   :  0.0   Min.   :  0.0
##  Class :character   1st Qu.:1.000   1st Qu.: 31.0   1st Qu.:888.0
##  Mode  :character   Median :2.000   Median : 76.0   Median :888.0
##                     Mean   :2.338   Mean   :383.2   Mean   :716.4
##                     3rd Qu.:3.000   3rd Qu.:888.0   3rd Qu.:888.0
##                     Max.   :8.000   Max.   :888.0   Max.   :888.0
##      ilp_r           mrace15          mhisp_r           mager9
##  Min.   :  0.0   Min.   : 1.000   Min.   :0.0000   Min.   :1.000
##  1st Qu.: 25.0   1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:3.000
##  Median : 57.0   Median : 1.000   Median :0.0000   Median :4.000
##  Mean   :339.3   Mean   : 1.854   Mean   :0.5083   Mean   :4.253
##  3rd Qu.:888.0   3rd Qu.: 1.000   3rd Qu.:0.0000   3rd Qu.:5.000
##  Max.   :888.0   Max.   :15.000   Max.   :5.0000   Max.   :9.000
##       dmar            meduc             wic              mhtr
##  Min.   :1.000   Min.   :0.000   Min.   :0.0000   Min.   : 0.00
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.:0.0000   1st Qu.:62.00
##  Median :1.000   Median :4.000   Median :0.0000   Median :64.00
##  Mean   :1.395   Mean   :4.203   Mean   :0.4038   Mean   :62.86
##  3rd Qu.:2.000   3rd Qu.:6.000   3rd Qu.:1.0000   3rd Qu.:66.00
##  Max.   :2.000   Max.   :8.000   Max.   :1.0000   Max.   :78.00
##       bmi             dwgt_r         wtgain          rf_pdiab
##  Min.   : 0.00   Min.   :  0   Min.   : 0.00   Min.   :0.000000
##  1st Qu.:21.60   1st Qu.:156   1st Qu.:20.00   1st Qu.:0.000000
##  Median :24.90   Median :178   Median :30.00   Median :0.000000
##  Mean   :26.07   Mean   :183   Mean   :30.22   Mean   :0.007507
##  3rd Qu.:29.80   3rd Qu.:206   3rd Qu.:39.00   3rd Qu.:0.000000
##  Max.   :99.90   Max.   :400   Max.   :98.00   Max.   :1.000000
##     rf_gdiab          rf_phype            ip_gon            ip_syph
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000   Median :0.000000   Median :0.00000
##  Mean   :0.05455   Mean   :0.01499   Mean   :0.002489   Mean   :0.00072
```

```
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00000
##     ip_chlam          ip_hepb           ip_hepc          oegest_comb
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.000000   Min.   :17.00
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.000000   1st Qu.:38.00
##  Median :0.00000   Median :0.000000   Median :0.000000   Median :39.00
##  Mean   :0.01807   Mean   :0.001996   Mean   :0.003101   Mean   :38.55
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.000000   3rd Qu.:40.00
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.000000   Max.   :47.00
##       lbw               smoked
##  Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.00000   Median :0.00000
##  Mean   :0.07685   Mean   :0.07266
##  3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.00000
```




3e+06
2e+06
1e+06
0e+00
count
0.00
0.25
0.50
0.75
1.00
lbw