ARE 213 Problem Set 1

Nick Depsky, Will Gorman, Peter Worley

October 1, 2018

Import Data

```
#setwd("~/Dropbox/Berkeley_tings/Fall 2018/ARE213/Problem Sets/PS1")
#setwd("C:\\Users\\will-\\Desktop\\are213")
setwd("C:\\Users\\Will\\Desktop\\are213")
dat <- read.dta("ps1.dta")
```

1a - Fix Missing Values (Last 15 columns)

```
dat_drop <- dat %>% filter(herpes != 8 & tobacco != 9 & cigar != 99 & cigar6 != 6 &
                     alcohol != 9 & drink != 99 & drink5 != 5 & wgain != 99)
```

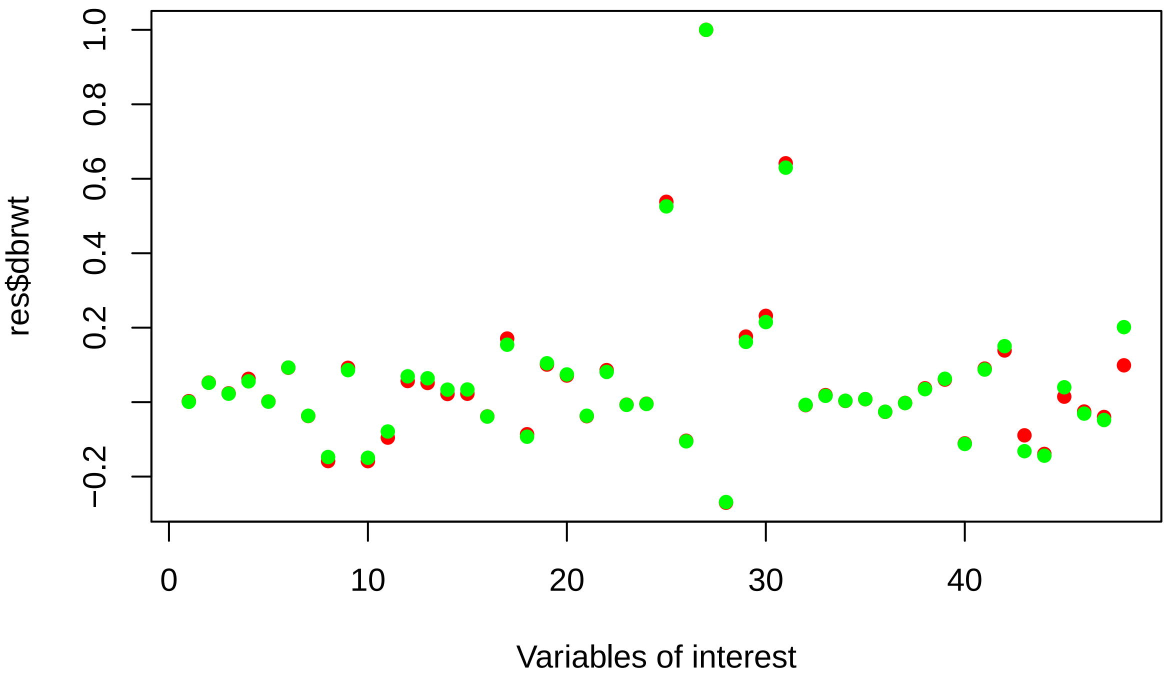
1b - Missing Data Discussion

The data being dropped were only from variables related to sexually transmitted disease (herpes), smoking and alcohol consumption, and weight gain. Each of these variables are more sensitive and potentially incriminating information for patient participants, and therefore may be underreported or undisclosed much more than other characteristics, such as having hypertension or anemia. The omission of such data therefore may not be random, but could be correlated with other variables that correlate with the incidence of these conditions and behaviors. Therefore, we might end up with lots of omitted individuals with high-risk lifestyles, which could produced biased results. One way to see if these data omissions truly are random would be to create dummy variables for each variable that has missing data (i.e. 0 - not missing, 1 - missing) and use this as an outcome variable in a logistic regression, to see if the beta coefficients of all other variables are statistically significant and non-zero. If so, we might conclude that the probability of a missing data value being present for a given high-risk lifestyle variable i not random, but in fact related to other non-missing variables.

Furthermore, we would likely want understand the data generating function for our outcome of interest (birth weight and APGAR score). We are mostly interested in variables within the dataset which would have an effect on the outcome. We would also be worried about variation within the missing data to see if the dropped data were significantly different with the remaining observational data of interest.

The data do not appear to be totally at random. In the below plot, we check the correlation values for some the variables both before (red) and after (green) dropping the observations. While most values don't change that noticeably, there are a few differences to question whether these observations were truly random.

```
#plot correlation matrix
res <- data.frame(cor(dat))
res2 <- data.frame(cor(dat_drop))
plot(res$dbrwt, pch = 16, col = "red",xlab = "Variables of interest")
points(res2$dbrwt, pch=16, col = "green")
```



1c - Summary Stats

```
sumstat <- as.data.frame(cbind(apply(dat_drop,2,mean),
                               apply(dat_drop,2,sd),
                               apply(dat_drop,2,min),
                               apply(dat_drop,2,max))) %>%
  set_colnames(c("Mean","SD","Min","Max")) %>% round(3)
```

	Variables	Mean	SD	Min	Max
rectype	Record Type	1.262	0.440	1	2
pldel3	Place of Birth Recode	1.018	0.133	1	2
birattnd	Attendant at Birth	1.202	0.564	1	5
cntocpop	Population of County of Occurence	1.443	1.137	0	3
stresfip	State of Residence (FIPS)	41.743	2.167	0	55
dmage	Age of Mother	27.757	5.699	12	49
ormoth	Hispanic Origin of Mother	0.091	0.522	0	5
mrace3	Race of Mother Recode	1.259	0.657	1	3
dmeduc	Education of Mother Detail	13.211	2.272	0	17
dmar	Marital Status of Mother	1.251	0.434	1	2
adequacy	Adequacy of Care Recode	1.297	0.546	1	3
nlbnl	Number of Live Births, Now Living	0.967	1.148	0	12
dlivord	Detail Live Birth Order	1.986	1.174	1	14
dtotord	Detail Total Birth Order	2.420	1.520	1	24

	Variables	Mean	SD	Min	Max
totord9	Total Birth Order Recode	2.407	1.458	1	8
monpre	Detail Month of Pregnancy Prenatal Care Began	2.502	1.326	0	9
nprevist	Total Number of Prenatal Visits	11.153	3.524	0	49
disllb	Interval Since Last Live Birth	350.412	362.325	0	777
isllb10	Interval Since Last Live Birth Recode	3.321	3.188	0	9
dfage	Age of Father	30.062	6.410	13	78
orfath	Hispanic Origin of Father	0.095	0.531	0	5
dfeduc	Education of Father Detail	13.277	2.325	0	17
birmon	Month of Birth	6.474	3.394	1	12
weekday	Day of Week of Birth	4.047	1.881	1	7
dgestat	Gestation - Detail in Weeks	39.153	2.445	17	47
csex	Sex	1.485	0.500	1	2
dbrwt	Birth Weight - Detail in Grams	3373.291	585.175	227	6067
dplural	Plurality	1.028	0.174	1	4
omaps	One Minute APGAR Score	8.117	1.260	0	10
fmaps	Five Minute APGAR Score	9.009	0.707	0	10
clingest	Clinical Estimate of Gestation	39.109	2.057	17	44
delmeth5	Method of Delivery Recode	1.549	1.010	1	5
anemia	Anemia	1.990	0.099	1	2
cardiac	Cardiac Disease	1.993	0.083	1	2
lung	Acute or Chronic Lung Disease	1.993	0.085	1	2
diabetes	Diabetes	1.973	0.162	1	2
herpes	Genital Herpes	1.994	0.078	1	2
chyper	Chronic Hypertension	1.992	0.087	1	2
phyper	Pregnancy-Associated Hypertension	1.969	0.172	1	2
pre4000	Previous Infant 4000+ Grams	1.986	0.119	1	2
preterm	Previous Preterm or Small-for-Gestational-Age Infant	1.986	0.118	1	2
tobacco	Tobacco Use During Pregnancy	1.841	0.366	1	2
cigar	Average Number of Cigarettes per Day	1.907	5.297	0	98
cigar6	Average Number of Cigarettes per Day Recode	0.346	0.861	0	5
alcohol	Alcohol Use During Pregnancy	1.990	0.098	1	2
drink	Average Number of Drinks per Week	0.031	0.619	0	91
drink5	Average Number of Drinks per Week Recode	0.020	0.230	0	4
wgain	Weight Gain in Pounds	30.356	11.884	0	98

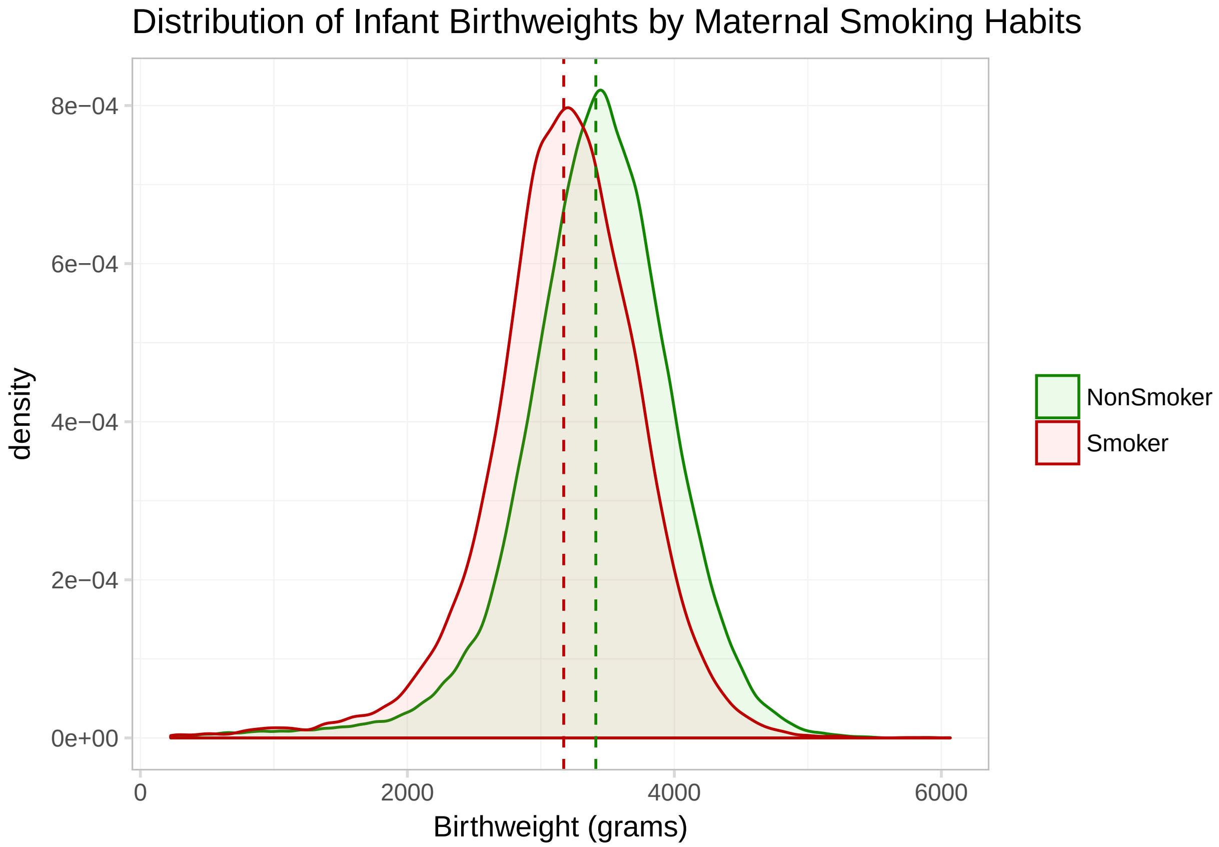
2a - Mean difference in APGAR scores

```r
keep.temp = c('omaps', 'fmaps', 'dbrwt')

group.ttest = function(x, group = as.factor(dat_drop$tobacco == 1) ){
  return(
    unlist(
      t.test( x ~ group)[c("estimate", "p.value")]
    )
  )
}

apgar = t(sapply(dat_drop[ , keep.temp], group.ttest))
colnames(apgar) <- c("non-smoker","smoker","p-value")
rownames(apgar) <- c("One-Minute APGAR Score", "Five-Minute APGAR Score", "Birthweight")
```

```r
stargazer(apgar, title = "Table 2a: different in apgar scores", type = "text")
```

```
##
## Table 2a: different in apgar scores
## ====================================================
##                         non-smoker  smoker   p-value
## ----------------------------------------------------
## One-Minute APGAR Score    8.120      8.103    0.088
## Five-Minute APGAR Score   9.009      9.009    0.979
## Birthweight             3,411.617  3,171.139    0
## ----------------------------------------------------
```

As we can see, only the difference in birthweight, and not APGAR scores, appears to be significant ($p < 0.05$) under the treatment of smoking while pregnant. The difference in birthweight distribution between smoking and non-smoking mothers during pregnancy is shown below.



2b - Average Treatment

One could identify the average treatment effect (ATE) of maternal smoking on birthweight by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers only if they were reasonably certain that all observable determinants of infant birth weight were measured for the sample, and any unobservable determinants of birth weight were accounted for via instrumental variables of some kind. Furthermore, the treatment, in this case smoking while pregnant, would have to be randomely distributed across all of these determinants, such that the likelihood of the treatment status of each individual was independent of the other determinants of birthweight. In other words, the treatment assignment is “as good as randomly assigned” after you condition on the observable factors, or other potential birthweight determinants.

If these assumptions were to hold, and we can claim that the difference in average birthweights of infants between mothers who were smokers during pregnancy and those that weren’t is in fact the ATE of smoking while pregnant, then we would calculte this ATE to be roughly -240.48 grams. In other words, we would claim that smoking while pregnant will, on average, reduce your child’s weight at birth by 240.48 grams.

```r
keep.temp = c('omaps', 'fmaps', 'dbrwt', "stresfip","dmage","ormoth","mrace3",
              "dmeduc","dmar","adequacy","dtotord","monpre","nprevist",
              "disllb","birmon","dgestat","csex","dplural","anemia","cardiac",
              "lung","diabetes","herpes","chyper","phyper","pre4000","preterm",
              "alcohol","drink","wgain")

group.ttest = function(x, group = as.factor(dat_drop$tobacco == 1) ){
  return(
    unlist(
      t.test( x ~ group)[c("estimate", "p.value")]
    )
  )
}

apgar = data.frame(t(sapply(dat_drop[ , keep.temp], group.ttest)))
colnames(apgar) <- c("non-smoker","smoker","p-value")

test <- data.frame(colMeans(dat_drop[ , keep.temp]))
colnames(test) <- c("pop avg")

final <- cbind(test,apgar)
final$treatment <- final$smoker - final$`non-smoker`
final <- as.matrix(final)

stargazer(final, title = "Table 2b: summary table of treatment effect on key variables", type = "text")
```

```
##
## Table 2b: summary table of treatment effect on key variables
## =========================================================
##           pop avg  non-smoker  smoker   p-value treatment
## ---------------------------------------------------------
## omaps      8.117     8.120      8.103    0.088   -0.017
## fmaps      9.009     9.009      9.009    0.979  -0.0001
## dbrwt    3,373.291 3,411.617  3,171.139    0    -240.478
## stresfip  41.743     41.720    41.865      0      0.145
## dmage     27.757     28.057    26.173      0     -1.883
## ormoth     0.091     0.096      0.064      0     -0.032
## mrace3     1.259     1.259      1.258    0.784   -0.001
## dmeduc    13.211     13.443    11.987      0     -1.456
## dmar       1.251     1.207      1.482      0      0.275
## adequacy   1.297     1.275      1.411      0      0.136
## dtotord    2.420     2.358      2.743      0      0.385
## monpre     2.502     2.454      2.754      0      0.300
## nprevist  11.153     11.252    10.626      0     -0.626
## disllb    350.412   358.713    306.633     0     -52.080
## birmon     6.474     6.473      6.481    0.794    0.007
## dgestat   39.153     39.173    39.047      0     -0.126
## csex       1.485     1.486      1.482    0.300   -0.004
## dplural    1.028     1.029      1.023   0.00000  -0.007
## anemia     1.990     1.991      1.986   0.00000  -0.004
## cardiac    1.993     1.993      1.994    0.135    0.001
## lung       1.993     1.993      1.990   0.0001   -0.003
## diabetes   1.973     1.973      1.973    0.984   0.00003
## herpes     1.994     1.994      1.993    0.320   -0.001
## chyper     1.992     1.992      1.993    0.040    0.001
## phyper     1.969     1.967      1.980      0      0.012
## pre4000    1.986     1.984      1.992      0      0.007
## preterm    1.986     1.988      1.975      0     -0.012
## alcohol    1.990     1.995      1.965      0     -0.030
## drink      0.031     0.011      0.136      0      0.125
## wgain     30.356    30.524     29.470      0     -1.054
## ---------------------------------------------------------
```

Based on Table 2b above, we might question some of the validity of the assumptions mentioned above. This table shows that there are other variables that are significantly correlated with the treatment of interest (i.e. smoking). Due to this correlation, we worry that the other variables of interest are not "as good as randomly assigned".

2c - Predetermination

Variables that could be considered predetermined are those that were determined prior to the treatment effect (i.e. prior to the current period being studied where mothers decision to smoke during the pregnancy). It is likely that these variables are things like the demographic information of the mother and father and the health information that do not correlate highly with smoking.

2d - Regression

Selection on observables implies that there is no selection into the treatment group due to unobserved characteristics of the observation. Another way of putting it is that we would be worried if there were additional characteristics (data) that we do not have access to (unobserved) that affect the outcome AND these characteristics are not randomly assigned. It is always a possibility that this is an issue. In our particular context, nutirition information is not included. For our below regression to work, we have to assume that nutrition qualities are as good as randomly assigned conditional on the data we do observe.

We analyzed the list of variables in the dataset, and decided on the following set of control variables to include in our regression, with birthweight as the outcome variable. We deemed these variables to be good controls to include in this regression due to the fact that there could be compelling arguments made for each of them as to why they may influence a mother's pregnancy and therein the health and birthweight of her child.

Table 2: Control Variables Included in Infant Birthweight Regression

Code	Variable
stresfip	State of Residence (FIPS)
dmage	Age of Mother
ormoth	Hispanic Origin of Mother
mrace3	Race of Mother Recode
dmeduc	Education of Mother Detail
dmar	Marital Status of Mother

Code	Variable
adequacy	Adequacy of Care Recode
dtotord	Detail Total Birth Order
monpre	Detail Month of Pregnancy Prenatal Care Began
nprevist	Total Number of Prenatal Visits
disllb	Interval Since Last Live Birth
birmon	Month of Birth
dgestat	Gestation - Detail in Weeks
csex	Sex
dplural	Plurality
anemia	Anemia
cardiac	Cardiac Disease
lung	Acute or Chronic Lung Disease
diabetes	Diabetes
herpes	Genital Herpes
chyper	Chronic Hypertension
phyper	Pregnancy-Associated Hypertension
pre4000	Previous Infant 4000+ Grams
preterm	Previous Preterm or Small-for-Gestational-Age Infant
tobacco	Tobacco Use During Pregnancy
cigar	Average Number of Cigarettes per Day
alcohol	Alcohol Use During Pregnancy
drink	Average Number of Drinks per Week
wgain	Weight Gain in Pounds

```r
lm.out <- lm(dbrwt ~ stresfip+dmage+ormoth+mrace3+dmeduc+dmar+adequacy+
               dtotord+monpre+nprevist+disllb+birmon+dgestat+csex+dplural+
               anemia+cardiac+lung+diabetes+herpes+chyper+phyper+pre4000+
               preterm+tobacco+cigar+alcohol+drink+wgain, data = dat_drop)
summary(lm.out)
```

```
## 
## Call:
## lm(formula = dbrwt ~ stresfip + dmage + ormoth + mrace3 + dmeduc + 
##     dmar + adequacy + dtotord + monpre + nprevist + disllb + 
##     birmon + dgestat + csex + dplural + anemia + cardiac + lung + 
##     diabetes + herpes + chyper + phyper + pre4000 + preterm + 
##     tobacco + cigar + alcohol + drink + wgain, data = dat_drop)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -2950.35  -289.55    -5.44   284.12  2738.96 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -5.467e+02  9.194e+01  -5.947 2.74e-09 ***
## stresfip     9.713e-01  6.152e-01   1.579  0.11435    
## dmage        1.820e+00  3.131e-01   5.814 6.12e-09 ***
## ormoth      -2.770e+01  2.603e+00 -10.638  < 2e-16 ***
## mrace3      -6.845e+01  2.279e+00 -30.035  < 2e-16 ***
## dmeduc       4.791e+00  7.024e-01   6.821 9.09e-12 ***
## dmar        -4.858e+01  3.938e+00 -12.337  < 2e-16 ***
## adequacy     1.134e+01  4.003e+00   2.833  0.00461 ** 
## dtotord      8.880e+00  1.181e+00   7.516 5.67e-14 ***
## monpre       1.744e+00  1.372e+00   1.272  0.20352
## nprevist     9.291e+00  5.195e-01  17.884  < 2e-16 ***
## disllb      -1.903e-01  4.715e-03 -40.372  < 2e-16 ***
## birmon      -3.811e-01  3.923e-01  -0.971  0.33131
## dgestat      1.060e+02  5.866e-01 180.780  < 2e-16 ***
## csex        -1.329e+02  2.665e+00 -49.860  < 2e-16 ***
## dplural     -6.439e+02  7.958e+00 -80.912  < 2e-16 ***
## anemia      -2.827e+01  1.343e+01  -2.105  0.03526 *
## cardiac      1.512e+01  1.611e+01   0.939  0.34797
## lung         1.478e+01  1.573e+01   0.940  0.34745
## diabetes    -1.626e+02  8.319e+00 -19.546  < 2e-16 ***
## herpes      -1.505e+01  1.705e+01  -0.883  0.37748
## chyper       1.339e+02  1.530e+01   8.749  < 2e-16 ***
## phyper       1.259e+02  7.783e+00  16.173  < 2e-16 ***
## pre4000     -3.779e+02  1.124e+01 -33.631  < 2e-16 ***
## preterm      2.508e+02  1.140e+01  22.005  < 2e-16 ***
## tobacco      1.591e+02  6.562e+00  24.246  < 2e-16 ***
## cigar       -3.804e+00  4.493e-01  -8.466  < 2e-16 ***
## alcohol      4.721e+01  1.583e+01   2.982  0.00287 **
## drink       -2.385e-01  2.491e+00  -0.096  0.92373
## wgain        8.340e+00  1.157e-01  72.074  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 450.4 on 114580 degrees of freedom
## Multiple R-squared:  0.4078, Adjusted R-squared:  0.4077
## F-statistic:  2721 on 29 and 114580 DF,  p-value: < 2.2e-16
```

With these results we can see that with multiple other potential influential variables controlled-for, the ATE of tobacco on birthweight looks to be about -159.1 grams, which is less than the -240.5 grams estimated from just looking at the difference in average birthweight between smokers and non-smokers only. Note that maternal smoking was coded as 2 for no smokers and 1 for smokers, so a positive beta coefficient for ‘tobacco’ implies higher birthweights for nonsmokers. Controlling for these additional variables suggests that the original ATE was like biased high by about 81.4 grams. This is due to the fact that mothers who smoke during pregnancy are also more likely to engage in behaviors or have other predetermined factors which negatively influence birthweight. Therefore, the original ATE estimate suffered from not controlling for other factors that were not randomly assigned.