ARE 213 Problem Set 1

```
rm(list = ls())
library(pacman)
p_load("foreign","dplyr","magrittr","knitr","myFuncs","ggplot2")
```

```
## Installing package into 'C:/Users/will-/Documents/R/win-library/3.4'
## (as 'lib' is unspecified)
```

```
## Warning: package 'myFuncs' is not available (for R version 3.4.3)
```

```
## Bioconductor version 3.6 (BiocInstaller 1.28.0), ?biocLite for help
```

```
## A new version of Bioconductor is available after installing the most
##   recent version of R; see http://bioconductor.org/install
```

```
## Warning in p_install(package, character.only = TRUE, ...):
```

```
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE,
## logical.return = TRUE, : there is no package called 'myFuncs'
```

```
## Warning in p_load("foreign", "dplyr", "magrittr", "knitr", "myFuncs", "ggplot2"): Failed to install/l
## myFuncs
```

```
theme_plot <- theme(
  legend.position = "right",
  panel.background = element_rect(fill = NA),
  panel.border = element_rect(fill = NA, color = "grey75"),
  axis.ticks = element_line(color = "grey85"),
  panel.grid.major = element_line(color = "grey95", size = 0.2),
  panel.grid.minor = element_line(color = "grey95", size = 0.2),
  legend.key = element_blank(),
  legend.title = element_blank(),
  legend.spacing.x = unit(0.3, "cm"))

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

1a - Fix Missing Values (Last 15 columns)

```
# Missing Data Codes (herpes = 8 (not stated in codebook), tobacco = 9, cigar = 99, cigar6 = 6, alcohol
dat %<>% filter(herpes != 8 & tobacco != 9 & cigar != 99 & cigar6 != 6 & alcohol != 9 & drink != 99 & di
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

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.

1c - Summary Stats

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

Variables <- c("Record Type","Place of Birth Recode","Attendant at Birth","Population of County of Occu
sumstat <- as.data.frame(cbind(Variables,sumstat))
kable(sumstat)
```

	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
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
dislb	Interval Since Last Live Birth	350.412	362.325	0	777
islb10	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

	Variables	Mean	SD	Min	Max
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

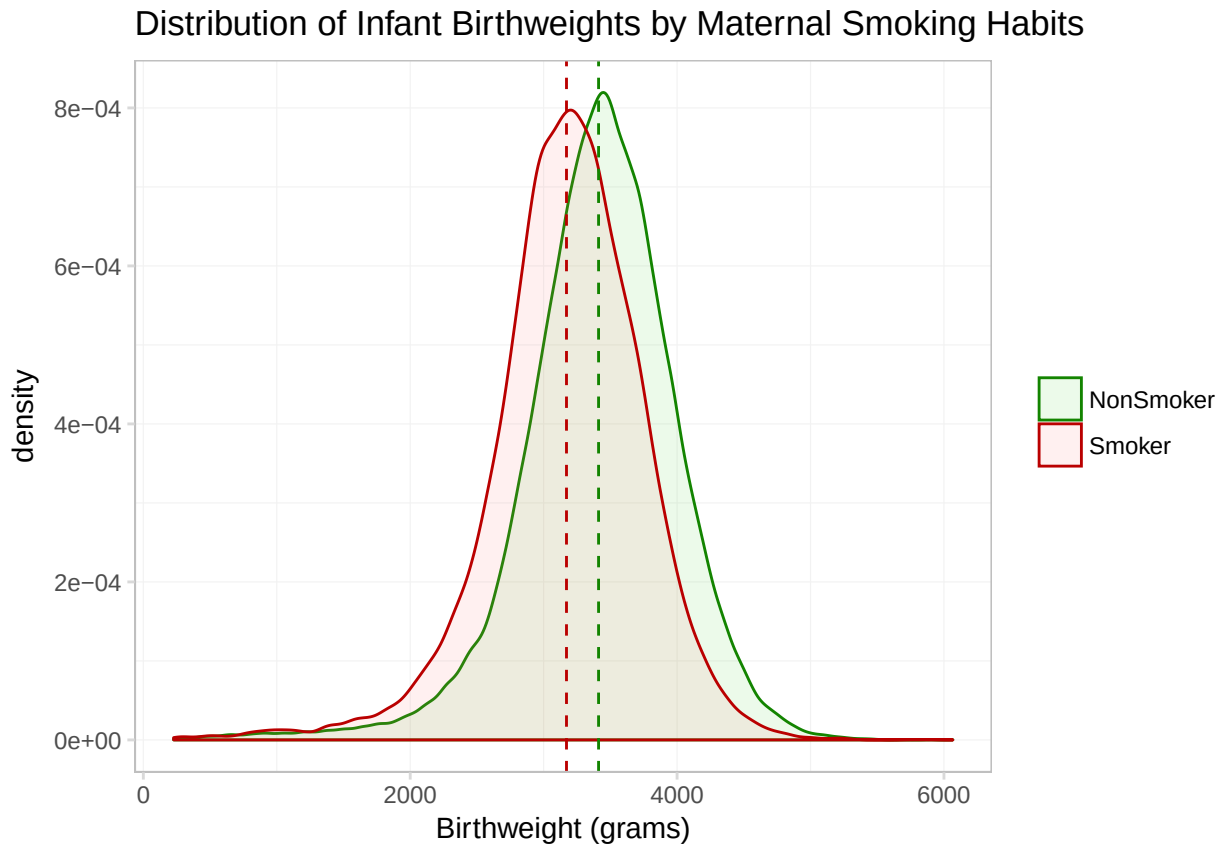
```
omaps <- cbind(c(mean(dat$omaps[dat$tobacco == 1]),mean(dat$omaps[dat$tobacco == 2])))
fmaps <- cbind(c(mean(dat$fmaps[dat$tobacco == 1]),mean(dat$fmaps[dat$tobacco == 2])))
bweight <- cbind(c(mean(dat$dbrwt[dat$tobacco == 1]),mean(dat$dbrwt[dat$tobacco == 2])))

psmoking.dat <- as.data.frame(cbind(omaps,fmaps,bweight)) %>% set_colnames(c("One-Minute APGAR Score",
```

2b

One could identify the ATE of maternal smoking 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 onobservable 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 infant birthweight. In other words, the treatment assignment is "as good as randomely 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 average treatment effect (ATE) of smoking while pregnant, then we would calculte this ATE to be roughly -240 grams. In other words, we would claim that smoking while pregnant will, on average, reduce your child's weight at birth by 240 grams.

```
ggdat <- dat %>% mutate(tobaccofact = dat$tobacco)
ggdat$tobaccofact[ggdat$tobaccofact == 1] <- "Smoker"
ggdat$tobaccofact[ggdat$tobaccofact == 2] <- "NonSmoker"
ggplot(ggdat, aes(dbrwt, colour = tobaccofact, fill = tobaccofact)) + geom_density(alpha = 0.1) + theme
```




```
psmoke.all <- dat %>% group_by(tobacco) %>% summarise_all(mean) %>% round(3) %>% t()
psmoke.all <- psmoke.all[-1,]
psmoke.all <- cbind(round(colMeans(dat[,-which(names(dat)=="tobacco")]),3),psmoke.all) %>% set_colnames
```