ARE 213 Problem Set 1A

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

Due 09/25/2020

Section 1

1. *Before getting started with the data work, first consider the table from Snow (1855) reproduced in the lecture notes ("Snow's Table IX"). The table reports only means.

(a) Develop an approximate 95% confidence interval for "Deaths per 10,000 Houses" for Southwark and Vaxhall customers. Develop another 95% CI for the same quantity for Lambeth. Do the confidence intervals overlap?

Note that that we're estimating p for a binomial distribution since deaths per 10,000 houses is the same as deaths per person (of course, scaled by persons per 10,000 households). Are we really dealing with a binomial distribution? Probably not, but it might not be a bad approximation if we think contaminated water is distributed randomly across space-time (so one person's probability exposure and subsequent death is the same and independent of another person's). Also, not everyone is equally susceptible to the virus (some have a higher p than others), but our estimate of p can be interpreted as an average p.

There are various ways to construct a confidence interval for an estimated binomial distribution. We use three different methods, all of which provide very similar estimates. The confidence intervals do not overlap.

```
# Southwark and Vauxhall
binom.confint(1263, 40046, method=c("asymptotic", "wilson", "agresti-coull"), type="central")
```

```
##          method    x     n       mean      lower      upper
## 1 agresti-coull 1263 40046 0.03153873 0.02987085 0.03329648
## 2    asymptotic 1263 40046 0.03153873 0.02982701 0.03325045
## 3        wilson 1263 40046 0.03153873 0.02987144 0.03329589
```

```
# Lambeth
binom.confint(98, 26107, method=c("asymptotic", "wilson", "agresti-coull"), type="central")
```

```
##          method  x     n        mean       lower       upper
## 1 agresti-coull 98 26107 0.003753783 0.003077893 0.004575688
## 2    asymptotic 98 26107 0.003753783 0.003011981 0.004495584
## 3        wilson 98 26107 0.003753783 0.003081460 0.004572122
```

(b) Discuss either formally or intuitively the critical assumption that underlies your confidence intervals. Give a 2 or 3 sentence quote from Snow's description (reproduced in Freedman (1991)) that supports this assumption.

To be confident that it is the choice of water company that is causing the difference in p and not some other factor, we need to be sure that there are not systematic differences between those who get their water from Southwark and Vauxhall and those who get it from Lambeth. John Snow argues that the two groups of people are comparable: "both rich and poor, both large houses and small" etc. In that case, we are reasonably certain that the difference in water company is what causes the difference in mortality risk.

Section 2

We now move to some analysis of real data. The data portions of Problem Sets 1a and 1b are based heavily on the paper Almond, Chay, and Lee (2005), and problem sets from Ken Chay and John DiNardo based on some of the data used in the paper. The goal of this assignment is to examine the research question: what is the causal effect of maternal smoking during pregnancy on infant birthweight and other infant health outcomes. The data for the problem set is an extract of all births from the 1993 National Natality Detail Files for Pennsylvania. Each observation represents an infant-mother match. The data in Stata format can be downloaded from the bCourses website. There should be 48 variables in the data and, after you are finished with the cleaning steps desribed below, 114,610 observations.

The data here are "real" and quite imperfect, which will help simulate the unpleasantness of real world data work. Unlike the real world where you will confront this bleak situation largely alone, I will provide you with some hints for working your way through the raw data. You can download part of the codebook for the data to help you figure out the relevant variables.

2. The first order of business is to go through the code book, decide on the relevant variables, and process the data. This involves several steps:

(a) Fix missing values. In the data set several variables take on a value of, say, 9999 if missing. We have already checked for missing observations for about 2/3 of the variables. The remaining variables need to be checked and are the last 15 in the variables list (i.e. from 'cardiac' to 'wgain'). Refer to the codebook for missing value codes. Produce an analysis data set that drops any observations with missing values.

```r
# According to the codebook, for the following medical risk factor variables, 8 corresponds to
# "Factor not on certificate" and 9 corresponds to "Factor not classifiable": cardiac, lung, diabetes, herpes

med_risk_factors <- c('cardiac', 'lung', 'diabetes', 'herpes', 'chyper', 'phyper', 'pre4000', 'preterm')

for (var in med_risk_factors){
  mom_dt[var] <- replace(mom_dt[var], which(mom_dt[var] == 8, arr.ind = TRUE), NA)
  mom_dt[var] <- replace(mom_dt[var], which(mom_dt[var] == 9, arr.ind = TRUE), NA)
}

# Below, arr.ind = TRUE returns the indices at which the row equals a certain value

# According to the codebook, for tobacco, 9 corresponds to "Unknown or not stated"
mom_dt$tobacco <- replace(mom_dt$tobacco, which(mom_dt$tobacco == 9, arr.ind = TRUE), NA)

# According to the codebook, for cigar, 99 corresponds to "Unknown or not stated"
mom_dt$cigar <- replace(mom_dt$cigar, which(mom_dt$cigar == 99, arr.ind = TRUE), NA)

# According to the codebook, for cigar6, 6 corresponds to "Unknown or not stated"
mom_dt$cigar6 <- replace(mom_dt$cigar6, which(mom_dt$cigar6 == 6, arr.ind = TRUE), NA)

# According to the codebook, for alcohol, 9 corresponds to "Unknown or not stated"
mom_dt$alcohol <- replace(mom_dt$alcohol, which(mom_dt$alcohol == 9, arr.ind = TRUE), NA)

# According to the codebook, for drink, 99 corresponds to "Unknown or not stated"
mom_dt$drink <- replace(mom_dt$drink, which(mom_dt$drink == 99, arr.ind = TRUE), NA)

# According to the codebook, for drink5, 5 corresponds to "Unknown or not stated"
mom_dt$drink5 <- replace(mom_dt$drink5, which(mom_dt$drink5 == 5, arr.ind = TRUE), NA)

# According to the codebook, for wgain (assuming that's wtgain in codebook),
# 99 corresponds to "Unknown or not stated"
mom_dt$wgain <- replace(mom_dt$wgain, which(mom_dt$wgain == 99, arr.ind = TRUE), NA)

# Make indicator for missing, will drop after comparison
setDT(mom_dt)
mom_dt[, miss := ifelse(complete.cases(mom_dt), 0, 1)]

# Now mom_dt contains 114,610 observations instead of the original 120,461
```

(b) If this were a real research project you would want to consider other approaches to missing data besides termination with extreme prejudice. What observations do you have to drop because of missing data? Might this affect your results? Do the data appear to be missing completely at random? How might you assess whether the data appear to be missing at random?

We know from the last problem set that if the data are missing at random then dropping them should not affect our results of the effect of smoking on birth weight. However, if the missing data is correlated with the treatment (smoking) or the outcome (birth weight) then it could bias our results. In the table below, we present the means and standard deviations for a number of the variables between the missing and nonmissing group, the difference in the means, t-statistic, and p-value (under the null that the difference in means is 0). From the table, it does appear that there are differences in the missing and nonmissing data. For example, the mothers in the missing data are younger, less educated, less likely to be married, have more previous children, received less prenatal care, have a shorter time since the last birth, and have a lower gestation period. As discussed in the last problem set, we could formally assess whether the data is missing at random by regressing an indicator for the missing variable on the treatment.

```r
# Compare missing to non missing
compare_dt <- transpose(mom_dt[,lapply(.SD, mean, na.rm=TRUE), .SDcols = c(6, 9:18, 20, 22, 25:30, 33:43, 45:
compare_dt <- cbind(compare_dt, transpose(mom_dt[,lapply(.SD, sd, na.rm=TRUE), .SDcols = c(6, 9:18, 20, 22, 2
colnames(compare_dt) <- c("Nonmiss means", "Miss means", "Nonmiss sd", "Miss sd")
compare_dt <- compare_dt[2:34,]
compare_dt[, Variable := c("Mother age", "Mother educ", "Marital status", "Prenatal adequacy", "Number living
           "Number dead or living child", "Total live birth or terminations", "Birth order", "Month prenatal
           "Number prenatal visits", "Time since last birth", "Father age", "Father educ", "Gestation", "Chil
           "Birth weight", "Number born", "One min Apgar", "Five min Apgar", "Anemia", "Cardiac disease",
           "Lung disease", "Diabetes", "Herpes", "Chron. hypertension", "Preg. hypertension", "Previous heavy
           "Previous preterm", "Tobacco use", "Number cigarettes", "Alcohol use", "Number drinks", "Weight ga

formulas <- paste("mom_dt$", names(mom_dt)[c(6, 9:18, 20, 22, 25:30, 33:43, 45:46, 48)], "~ mom_dt$miss")
t_test <- t(sapply(formulas, function(f) {
  res <- t.test(as.formula(f))
  c(res$statistic, p.value=res$p.value)
}))

colnames(t_test) <- c("t-stat", "p-value")

compare_dt <- cbind(compare_dt, t_test)
compare_dt[, Difference := `Nonmiss means` - `Miss means`]

setcolorder(compare_dt, c("Variable", "Nonmiss means", "Nonmiss sd", "Miss means", "Miss sd", "Difference", "

mom_dt <- na.omit(mom_dt)
```

```r
print(xtable(compare_dt, caption = 'Difference in Means Missing v Nonmissing', digits = 2),
      include.rownames = FALSE, size = "small", comment = FALSE)
```

Variable	Nonmiss means	Nonmiss sd	Miss means	Miss sd	Difference	t-stat	p-value
Mother age	27.76	5.70	27.05	5.97	0.71	8.84	0.00
Mother educ	13.21	2.27	12.51	2.26	0.70	23.16	0.00
Marital status	1.25	0.43	1.44	0.50	-0.19	-28.00	0.00
Prenatal adequacy	1.30	0.55	1.63	0.79	-0.33	-31.58	0.00
Number living child	0.97	1.15	1.24	1.43	-0.27	-14.16	0.00
Number dead or living child	1.99	1.17	2.27	1.47	-0.28	-14.38	0.00
Total live birth or terminations	2.42	1.52	2.81	1.87	-0.39	-15.76	0.00
Birth order	2.41	1.46	2.78	1.74	-0.37	-15.96	0.00
Month prenatal began	2.50	1.33	2.80	1.92	-0.30	-11.79	0.00
Number prenatal visits	11.15	3.52	9.32	4.90	1.84	28.30	0.00
Time since last birth	350.41	362.33	315.97	355.26	34.44	7.23	0.00
Father age	30.06	6.41	29.61	7.04	0.46	4.84	0.00
Father educ	13.28	2.33	12.67	2.29	0.60	19.61	0.00
Gestation	39.15	2.44	38.53	3.42	0.62	13.77	0.00
Child sex	1.49	0.50	1.48	0.50	0.00	0.14	0.89
Birth weight	3373.29	585.17	3191.90	716.95	181.39	19.03	0.00
Number born	1.03	0.17	1.04	0.21	-0.01	-4.17	0.00
One min Apgar	8.12	1.26	7.90	1.57	0.21	10.18	0.00
Five min Apgar	9.01	0.71	8.88	1.03	0.13	9.47	0.00
Anemia	1.99	0.10	1.99	0.12	0.00	2.66	0.01
Cardiac disease	1.99	0.08	1.99	0.09	0.00	0.70	0.49
Lung disease	1.99	0.08	1.99	0.10	0.00	1.57	0.12
Diabetes	1.97	0.16	1.97	0.16	0.00	0.07	0.94
Herpes	1.99	0.08	1.99	0.10	0.00	2.44	0.01
Chron. hypertension	1.99	0.09	1.99	0.10	0.00	1.31	0.19
Preg. hypertension	1.97	0.17	1.97	0.16	-0.00	-2.03	0.04
Previous heavy birth	1.99	0.12	1.99	0.10	-0.00	-2.81	0.01
Previous preterm	1.99	0.12	1.98	0.16	0.01	5.16	0.00
Tobacco use	1.84	0.37	1.57	0.50	0.27	41.11	0.00
Number cigarettes	1.91	5.30	3.94	7.42	-2.03	-18.74	0.00
Alcohol use	1.99	0.10	1.63	0.48	0.36	56.48	0.00
Number drinks	0.03	0.62	0.16	1.47	-0.13	-5.33	0.00
Weight gain	30.36	11.88	30.78	13.14	-0.43	-1.71	0.09

Table 1: Difference in Means Missing v Nonmissing

(c) Produce a summary table describing the final analysis data set.

We create a summary table similar to the table in b, but this time we compare the means of smokers vs non-smokers. To see means/standard deviations for the entire dataset, refer to the nonmissing data columns of part b. We will discuss the differences between the groups in 3b.

```r
# Recode to binary 0/1 treatment
# tobacco is 1: yes, tobacco use during pregnancy and 2: no tobacco use during pregnancy
mom_dt[, tobacco := ifelse(tobacco==2, 0, 1)]

# Compare smoker to nonsmoker
summary_dt <- transpose(mom_dt[,lapply(.SD, mean, na.rm=TRUE), .SDcols = c(6, 9:18, 20, 22, 25:30, 33:41, 43,
summary_dt <- cbind(summary_dt, transpose(mom_dt[,lapply(.SD, sd, na.rm=TRUE), .SDcols = c(6, 9:18, 20, 22, 2
colnames(summary_dt) <- c("Nonsmoker means", "Smoker means", "Nonsmoker sd", "Smoker sd")
summary_dt <- summary_dt[2:33,]
summary_dt[, Variable := c("Mother age", "Mother educ", "Marital status", "Prenatal adequacy", "Number living
           "Number dead or living child", "Total live birth or terminations", "Birth order", "Month prenatal
           "Number prenatal visits", "Time since last birth", "Father age", "Father educ", "Gestation", "Chil
           "Birth weight", "Number born", "One min Apgar", "Five min Apgar", "Anemia", "Cardiac disease",
           "Lung disease", "Diabetes", "Herpes", "Chron. hypertension", "Preg. hypertension", "Previous heavy
           "Previous preterm", "Number cigarettes", "Alcohol use", "Number drinks", "Weight gain")]

formulas <- paste("mom_dt$", names(mom_dt)[c(6, 9:18, 20, 22, 25:30, 33:41, 43, 45:46, 48)], "~ mom_dt$tobacc
t_test <- t(sapply(formulas, function(f) {
  res <- t.test(as.formula(f))
  c(res$statistic, p.value=res$p.value)
}))

colnames(t_test) <- c("t-stat", "p-value")

summary_dt <- cbind(summary_dt, t_test)
summary_dt[, Difference := `Nonsmoker means` - `Smoker means`]

setcolorder(summary_dt, c("Variable", "Nonsmoker means", "Nonsmoker sd", "Smoker means", "Smoker sd", "Differ
```

```
print(xtable(summary_dt, caption = 'Difference in Means Smoker v Nonsmoker', digits = 2),
      include.rownames = FALSE, size = "small", comment = FALSE)
```

Variable	Nonsmoker means	Nonsmoker sd	Smoker means	Smoker sd	Difference	t-stat	p-value
Mother age	28.06	5.67	26.17	5.61	1.88	41.56	0.00
Mother educ	13.44	2.30	11.99	1.63	1.46	102.72	0.00
Marital status	1.21	0.41	1.48	0.50	-0.27	-70.08	0.00
Prenatal adequacy	1.28	0.53	1.41	0.63	-0.14	-27.41	0.00
Number living child	0.93	1.13	1.15	1.22	-0.22	-22.60	0.00
Number dead or living child	1.95	1.15	2.18	1.27	-0.23	-22.98	0.00
Total live birth or terminations	2.36	1.48	2.74	1.67	-0.39	-29.06	0.00
Birth order	2.35	1.42	2.73	1.60	-0.38	-29.95	0.00
Month prenatal began	2.45	1.28	2.75	1.51	-0.30	-25.11	0.00
Number prenatal visits	11.25	3.45	10.63	3.84	0.63	20.52	0.00
Time since last birth	358.71	364.07	306.63	349.76	52.08	18.33	0.00
Father age	30.27	6.34	28.96	6.65	1.31	24.60	0.00
Father educ	13.49	2.37	12.13	1.67	1.37	94.00	0.00
Gestation	39.17	2.39	39.05	2.71	0.13	5.88	0.00
Child sex	1.49	0.50	1.48	0.50	0.00	1.04	0.30
Birth weight	3411.62	579.73	3171.14	572.08	240.48	51.98	0.00
Number born	1.03	0.18	1.02	0.15	0.01	5.26	0.00
One min Apgar	8.12	1.26	8.10	1.27	0.02	1.71	0.09
Five min Apgar	9.01	0.71	9.01	0.71	0.00	0.03	0.98
Anemia	1.99	0.10	1.99	0.12	0.00	4.61	0.00
Cardiac disease	1.99	0.08	1.99	0.08	-0.00	-1.50	0.13
Lung disease	1.99	0.08	1.99	0.10	0.00	3.80	0.00
Diabetes	1.97	0.16	1.97	0.16	-0.00	-0.02	0.98
Herpes	1.99	0.08	1.99	0.08	0.00	0.99	0.32
Chron. hypertension	1.99	0.09	1.99	0.08	-0.00	-2.05	0.04
Preg. hypertension	1.97	0.18	1.98	0.14	-0.01	-10.51	0.00
Previous heavy birth	1.98	0.12	1.99	0.09	-0.01	-9.16	0.00
Previous preterm	1.99	0.11	1.98	0.15	0.01	10.37	0.00
Number cigarettes	0.00	0.00	11.96	7.47	-11.96	-216.57	0.00
Alcohol use	2.00	0.07	1.97	0.18	0.03	21.83	0.00
Number drinks	0.01	0.25	0.14	1.44	-0.13	-11.77	0.00
Weight gain	30.52	11.56	29.47	13.45	1.05	9.92	0.00

Table 2: Difference in Means Smoker v Nonsmoker

3. The next part of the assignment is to try to estimate the “causal” effect of maternal smoking during pregnancy on infant birth weight. Let’s start out using techniques that are familiar, and think about whether they are likely to work in this context. Answer the following questions.

(a) Compute the mean difference in APGAR scores (both five and one minute versions) as well as birthweight by smoking status.

```r
# According to the codebook, omaps is the one minute APGAR score and fmaps is the five minute APGAR score
# Both are a score from 0-10
# dbrwt (assuming that corresponds to dbirwt in codebook) is birthweight in grams

smoker <- subset(mom_dt, mom_dt$tobacco == 1)
nonsmoker <- subset(mom_dt, mom_dt$tobacco == 0)

# Mean difference in one minute APGAR score by smoking status
mean_diff_1min_apgar <- mean(smoker$omaps) - mean(nonsmoker$omaps)
print(mean_diff_1min_apgar)
```

```
## [1] -0.01743508
```

```r
# Mean difference in five minute APGAR score by smoking status
mean_diff_5min_apgar <- mean(smoker$fmaps) - mean(nonsmoker$fmaps)
print(mean_diff_5min_apgar)
```

```
## [1] -0.0001498085
```

```r
# Mean difference in birthweight by smoking status
mean_diff_birthweight <- mean(smoker$dbrwt) - mean(nonsmoker$dbrwt)
print(mean_diff_birthweight)
```

```
## [1] -240.4778
```

(b) Under what circumstances can one identify the average treatment effect of maternal smoking by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers? Estimate its impact under this assumption. Provide and comment on some evidence for or against the validity of the assumption.

We can identify the average treatment effect by comparing the unadjusted difference in means if the treatment (smoking) is randomly assigned. Under this assumption, we would estimate that smoking results in an effect of -240.478 grams (? check this) on birth weight. The mean of all birth weights is 3373.291 so this represents a decrease of -0.071 percent.

However, based on our table in 2c, random assignment is likely not a valid assumption. Smokers and non smokers are different on many dimensions, some of which may affect birth weight. For example, smokers started prenatal care later and had fewer prenatal visits.

(c) Suppose that maternal smoking is randomly assigned conditional on the other observable “predetermined” determinants of infant birth weight. First discuss which (if any) of the variables contained in the data set can clearly be considered to be predetermined. In general, what kinds of variables can be considered predetermined and what kinds of variables cannot?

A variable can be considered “predetermined" if it affects selection into treatment (smoking), and there is no reverse causality between the treatment and the predetermined variable. A variable that both affects selection into treatment, but may also be affected by treatment status, is not considered predetermined. In our data set, the variables on state of residence, characteristics of county of residence, age, race, education, and marital status of the mother and the father can be considered to be “predetermined.” In contrast, variables such as on prenatal care, alcohol use during pregnancy, weight gain during pregnancy, and birth month (i.e. whether the birth is premature) would not be considered predetermined, since they may be affected by the treatment status.

(d) What does “selection on observables" imply about the relationship between maternal smoking and unobservable determinants of birth weight conditional on the observables? Use a basic linear regression model, in conjunction with your answer to (c), to estimate the impact of smoking and report your estimates. Under what circumstances is the average treatment effect identified?

The key assumption underlying a “selction on observables" design is that the treatment is as good as randomly assigned after we condition on observables. In other words, we assume that we observe *all* the factors that affect treatment assignment (smoking) and are correlated with the potential outcomes (birth weight). If there is systematic selection into treatment, we assume this selection is only a function of the observabless. That is, we assume that maternal smoking is uncorrelated with unobservable determinants of birth weight conditional on the observables. If these assumptions hold, then a regression of maternal smoking on birth weight, conditioning on observables, will estimate the ATE.

In a selection on observables design, we estimate the model:

$$birthweight_i = \alpha_i + \beta Smoking_i + \delta_1 mother_age_i + \delta_2 mother_educ_i + \delta_3 mother_race + \delta_4 marital_status_i + \delta_5 father_age_i + \delta_6 father_educ + \delta_7$$

```r
# Selection on observables model
lm1 <- lm(dbrwt ~ tobacco + dmage + dmeduc + mrace3 + dmar + dfage + dfeduc + stresfip + cntocpop, mom_dt)

summary(lm1)
```

```
##
## Call:
## lm(formula = dbrwt ~ tobacco + dmage + dmeduc + mrace3 + dmar +
##     dfage + dfeduc + stresfip + cntocpop, data = mom_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3252.3  -305.1    27.4   355.4  2850.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3384.2925    37.1691  91.051  < 2e-16 ***
## tobacco     -211.8008     4.8445 -43.720  < 2e-16 ***
## dmage          1.8369     0.5009   3.667 0.000245 ***
## dmeduc         2.0521     1.0313   1.990 0.046609 *
## mrace3       -98.4385     2.9267 -33.634  < 2e-16 ***
## dmar         -75.5380     4.9182 -15.359  < 2e-16 ***
## dfage          0.6108     0.4153   1.471 0.141380
## dfeduc         2.3924     0.9881   2.421 0.015476 *
## stresfip       2.2002     0.7801   2.820 0.004797 **
## cntocpop      14.6549     1.5773   9.291  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.4 on 114600 degrees of freedom
## Multiple R-squared:  0.0498, Adjusted R-squared:  0.04972
## F-statistic: 667.3 on 9 and 114600 DF,  p-value: < 2.2e-16
```

If our selection on observables assumption holds, we can interpret these results as indicating that maternal smoking causes a decrease in birth weight of 211.8 grams (a decrease of 0.063 percent from a base of 3,384.3, which is statistically significant at the 1% level.