Statistics 641, Fall 2013
Homework #1
Solutions

1. Suppose in a randomized trial, subjects' responses are measured at three times: baseline (randomization), month 1, month 2. The data file "data1.csv" (in csv format, comma delimited) contains columns

   - `x0`: baseline value of response variable
   - `x1`: value of response variable at month 1
   - `x2`: value of response variable at month 2
   - `z`: treatment variable (0,1).

   Note: this file can be read into `R` using the command

   ```
   > data = read.csv(data1.csv)
   ```

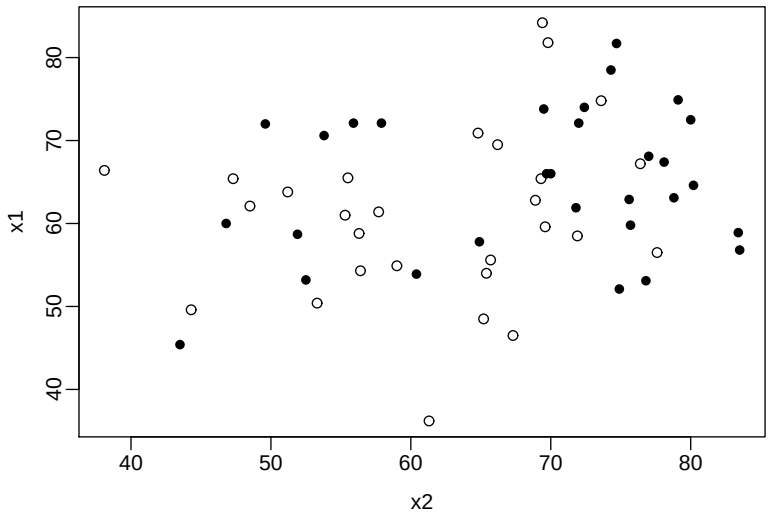
   Assume that the responses are normally distributed.

   (a) For the response at month 1 (`x1`), compute the mean and standard deviation for each treatment group ($z = 0, 1$)

   ```
   > data <- read.csv("../hw/hw/data1.csv")
   > aggregate(x1~z,mean,data=data)
     z       x1
   1 0 60.91429
   2 1 64.80000
   > aggregate(x1~z,sd,data=data)
     z        x1
   1 0 10.341957
   2 1  8.792709
   ```

   (b) Plot the month 1 response against the month 2 response, using a different plotting symbol for each treatment group.

   ```
   > plot(x1~x2, data=data, pch=c(1,16)[z+1])
   ```



(c) Perform a two-sample t-test comparing the month 1 response between treatment groups (you may assume equal variances).

```
> summary(lm(x1~z, data=data))
 ...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   60.914      1.809  33.677   <2e-16 ***
z              3.886      2.515   1.545    0.128
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The t-statistic is 1.545.

(d) Fit a linear model comparing the month 1 response by treatment adjusted for the baseline value, `x0`.

```
> summary(lm(x1~z+x0, data=data))
...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.7465     3.6149   3.803 0.000361 ***
z             3.6458     1.2252   2.976 0.004337 **
x0            0.7709     0.0573  13.454  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.662 on 55 degrees of freedom
F-statistic: 95.54 on 2 and 55 DF,  p-value: < 2.2e-16
```

Adjusted for baseline, the t-statistic is 2.976

(e) Why does (d) yield a different answer than (c)? Is `x0` a confounder?

The t-statistic is much larger in (d) than in (c). In (c), the mean difference is 3.886, and in (d) it is 3.646, whereas in (c) the standard error of the difference is 2.515 versus 1.225 in (d). Accounting for the association between baseline and month 1 increases the precision of the estimate of difference.

Regardless of this result, `x0` cannot be confounder because the study is randomized.

(f) Perform a two-sample t-test comparing the month 2 response between treatment groups.

```
> summary(lm(x2~z, data=data))
...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.618      2.063  29.868   <2e-16 ***
z              6.872      2.868   2.396     0.02 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.92 on 56 degrees of freedom
```

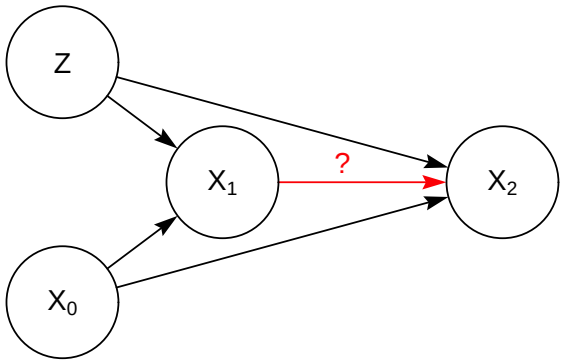
(g) Fit a linear model comparing the month 2 response by treatment adjusted for the month 1 value, `x1`.

```
> summary(lm(x2~z+x1, data=data))
...

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  47.6754     9.4015   5.071 0.00000481 ***
z             5.9828     2.8954   2.066     0.0435 *
x1            0.2289     0.1507   1.519     0.1344
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.79 on 55 degrees of freedom
```

(h) Why does (g) yield a different answer than (f)? Is `x1` a confounder?

The t-statistic is larger in (g) than in (f) because `x1` accounts for some (small amount) of the effect of treatment on `x2`. However, because the study is randomized `x1` cannot be a confounder. In fact, `x1` is observed *after* treatment start, so the causal pathway is from `z` to `x1`, rather than `x1` to `z`, which is required for confounding.

Consider the diagram:



In the model

$$X_2 = \alpha + \beta Z + \epsilon$$

randomization ensures that β is the true treatment effect. So

$$E[X_2|Z=1] - E[X_2|Z=0] = \beta \tag{1}$$

On the other hand given the model:

$$X_2 = \alpha + \beta' Z + \gamma X_1 + \epsilon$$

Average across each treatment group:

$$\begin{aligned} E[X_2|Z=1] &= \alpha + \beta' + \gamma E[X_1|Z=1] \\ E[X_2|Z=0] &= \alpha + \gamma E[X_1|Z=0] \end{aligned}$$

Average difference between treatment groups:

$$E[X_2|Z=1] - E[X_2|Z=0] = \beta' + \gamma\big(E[X_1|Z=1] - E[X_1|Z=0]\big)$$

This is the same quantity as (1) above, so

$$\beta' = \beta - \gamma\big(E[X_1|Z=1] - E[X_1|Z=0]\big)$$

Conclusion: β' is "contaminated" by "adjustment for X_1" and doesn't represent the real effect of treatment. Unless you know what you are doing, *never* adjust for post-randomization variables.