**Statistics 641, Fall 2014**
**Homework #1**
**Solutions**

1. Suppose in a randomized trial, we observe a baseline (pre-randomization) variable, `w`, and two response variables, `x` and `y`. The data file "data1.csv" (in csv format, comma delimited) contains columns

   - `z`: binary treatment variable (0,1).
   - `w`: baseline variable
   - `x`: response variable
   - `y`: response variable

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

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

   Assume that the all variables other than `z` are normally distributed. (Note that I've deleted lots of extraneous output in what follows.)

   (a) Perform a two-sample $t$-test for differences in variable `w` between treatment groups (you may assume equal variances). What does this analysis tell you?

   ```
   > data <- read.csv("data1.csv")
   > summary(lm(w~z, data=Y))
   ...
   Coefficients:
                Estimate Std. Error t value Pr(>|t|)
   (Intercept)   3.0120     0.1497  20.118   <2e-16 ***
   z            -0.0040     0.2117  -0.019    0.985
   ```

   The difference in group means is very small, and the $p$-value quite large (0.985), however, because of the randomization, any difference between groups is necessarily due to chance. Hence, the $p$-value in particular has no useful interpretation.

   (b) For response variable `x`, compute the mean and standard deviation for each treatment group ($z = 0, 1$)

   ```
   > aggregate(x~z,mean,data=data)
     z      x
   1 0 13.944
   2 1 14.336
   > aggregate(x~z,sd,data=data)
     z         x
   1 0 1.0913687
   2 1 0.9855611
   ```

(c) Perform a two-sample $t$-test comparing response `x` between treatment groups (assume equal variances).

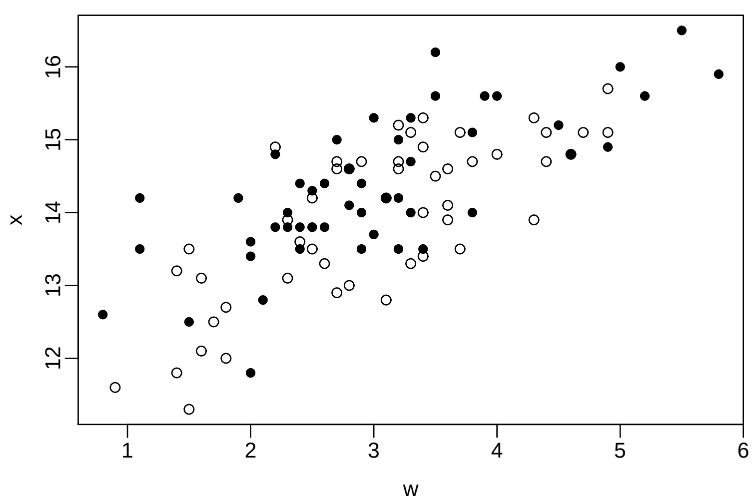
```
> summary(lm(x~z, data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.9440     0.1471  94.824   <2e-16 ***
z             0.3920     0.2080   1.885   0.0624 .
...
Residual standard error: 1.04 on 98 degrees of freedom
```

The $t$-statistic is 1.885, which does not reach traditional levels of statistical significance.

(d) Plot `x` versus `w`, using a different plotting symbol for each treatment group.

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




(e) Fit a linear model comparing response `x` by treatment adjusted for the baseline value, `w`.

```
> summary(lm(x~z+w, data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.73170    0.22227  52.781  < 2e-16 ***
z            0.39494    0.13878   2.846  0.00541 **
w            0.73449    0.06621  11.093  < 2e-16 ***

...

Residual standard error: 0.6945 on 97 degrees of freedom
```

Adjusted for baseline, the $t$-statistic is 2.846, which does reach traditional levels of statistical significance. Note further that with a $t$-statistic of 11.1, `w` is strongly associated with `x`.

(f) Why does (e) yield a different answer than (c)? Is `w` a confounder?

The $t$-statistic is larger in (e) than in (c), however, in (c) the mean difference is 0.392, and in (e) the mean difference is 0.3949, a negligible difference, whereas in (c) the standard error of the difference is 0.2080 versus 0.1388 in (e). Furthermore, the smaller residual standard error in (e) indicates that `w` accounts for a meaningful amount of the variability in `x` and thereby increases the precision of the estimate of difference.

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

(g) Perform a two-sample $t$-test comparing the response `y` between treatment groups (again assume equal variances).

```
> summary(lm(y~z, data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   49.980      1.390  35.949   <2e-16 ***
z              3.320      1.966   1.689   0.0945 .

...

Residual standard error: 9.831 on 98 degrees of freedom
```

The $t$-statistic is 1.689.

(h) Perform a two-sample $t$-test comparing the response `y` between treatment groups, adjusted for `w`.

```
> summary(lm(y~z+w, data=data))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.6902     2.5730  13.094  < 2e-16 ***
z             3.3416     1.6066   2.080   0.0402 *
w             5.4083     0.7665   7.056 2.56e-10 ***

...

Residual standard error: 8.033 on 97 degrees of freedom
```

The $t$-statistic adjusted for `w` is 2.080.

(i) Fit a linear model comparing response `y` by treatment adjusted for both `w` and `x`.

```
> summary(lm(y~z+w+x, data=data))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -63.62001    9.83542  -6.468 4.15e-09 ***
z             0.06577    1.17256   0.056    0.955
w            -0.68406    0.80948  -0.845    0.400
x             8.29463    0.82414  10.065  < 2e-16 ***

...

Residual standard error: 5.632 on 96 degrees of freedom
```

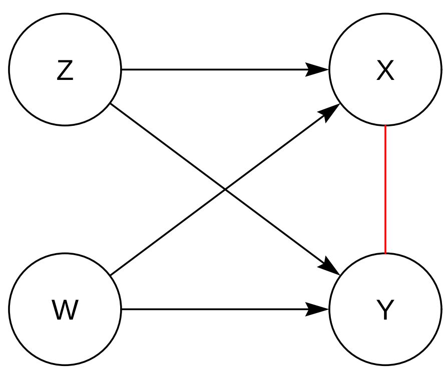
The $t$-statistic adjusted for both `w` and `x` is 0.056.

(j) Compare the results from (g), (h), and (i). Is `x` a confounder for the association between `z` and `y`?

As in (f), the $t$-statistic is larger in (h) than in (g) because `w` accounts for some of the variability of treatment on `y`: the point estimates are quite similar, but the standard error in (h) is smaller than in (g).

Further "adjustment" for `x` makes the association between `z` and `y` disappear, however, because the study is randomized `x` cannot be a confounder. In fact, `x` is observed *after* treatment start, so the causal pathway is from `z` to `x`, rather than `x` to `z`, which is required for confounding.

Consider the diagram:




$W$ and $Z$ are independent because $Z$ is randomly assigned. Both $X$ and $Y$ differ by treatment (based on analysis adjusted for $W$), and because of the randomization, this must be causal. Because $W$ precedes $X$ and $Y$ in time, I've drawn arrows from $W$ to each of these, although it is unclear that we can declare this relationship as "causal". Clearly $X$ and $Y$ are associated, but it is unclear in which direction the arrow should go, if any.

For the model

$$Y = \alpha + \beta Z + \delta W + \epsilon$$

randomization ensures that $\beta$ is the true treatment effect:

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

On the other hand given the model:

$$Y = \alpha + \beta' Z + \delta W + \gamma X + \epsilon$$

If we average within each treatment group, (by randomization, $W$ is independent of $Z$)

$$\begin{aligned} E[Y|Z=1] &= \alpha + \beta' + \delta E[W] + \gamma E[X|Z=1] \\ E[Y|Z=0] &= \alpha + \delta E[W] + \gamma E[X|Z=0], \end{aligned}$$

we see that the average difference between treatment groups is

$$E[Y|Z=1] - E[Y|Z=0] = \beta' + \gamma\big(E[X|Z=1] - E[X|Z=0]\big).$$

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

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

and because $X$ is related to $Z$, $\beta \neq \beta'$.

**Conclusion:** $\beta'$ is "contaminated" by "adjustment for $X$" and doesn't represent the actual effect of treatment on $Y$, which is $\beta$. Unless you know what you are doing, *never* adjust for post-randomization variables.