Confounding and Collinearity in Multiple Linear Regression

Basic Ideas

Confounding: A third variable, not the dependent (outcome) or main independent (exposure) variable of interest, that distorts the observed relationship between the exposure and outcome. Confounding complicates analyses owing to the presence of a third factor that is associated with both the putative risk factor and the outcome.

For example, consider the mortality rate in Florida, which is much higher than in Michigan. Before concluding that Florida is a riskier place to live, one needs to consider confounding factors such as age. Florida has a higher proportion of people of retirement age and older than does Michigan, and older people are more likely to die in any given interval of time. Therefore, one must "adjust" for age before drawing any conclusions.

One of the epidemiologist's tools for discovering and correcting confounding is stratification, which in the preceding example would have the epidemiologist compare mortality rates in Florida and Michigan separately for people in across a range of age groups. Indeed, such stratified analyses, should often be a first step towards investigating confounding.

Another way would be to use multiple regression, to derive mortality rates for Florida compared to Michigan adjusted for any differences in age (and possibly for other confounding factors).

Criteria for a confounding factor:

1. A confounder must be a risk factor (or protective factor) for the outcome of interest.
2. A confounder must be associated with the main independent variable of interest. For example, the confounder must be unevenly distributed as far as the independent variable is concerned, as in the smoking and occupation example of last class. Smoking was a confounder for the outcome of cancer, because smoking is associated with cancer, and was unevenly distributed among occupation categories.
3. A confounder must not be an intermediate step in the causal pathway between the exposure and outcome. [This last criterion above has been controversial of late.]

Confounding arises:

- Confounding by indication: When evaluating the effect of a particular drug, many times people who take the drug differ from those who do not according to the medical indication for which the drug is prescribed.
- Selection bias: Not everyone invited to participate in a study participates, causing imbalance between groups.
- Recall bias: Not everyone with an exposure recalls their exposure history correctly, perhaps causing uneven recall in different groups.
- Many other ways, including unbalanced groups by chance, especially in smaller studies.

Example:

Hypothesis: Caffeine intake is associated with heart disease. Which of the following are likely to be confounding factors?

Factor	Low Caffeine Intake	High Caffeine Intake
Current Smoker (%)	12%	27%
Age (mean years)	36.3	37.1
Body Mass Index (mean)	28.4	24.3
Regular Exercise (%)	24%	14%
Female Gender (%)	43%	41%
Type A personality (%)	16%	28%
Hypertension (%)	9%	16%

Any variables with imbalances are potential confounders, such as smoking, BMI, exercise, and Type A personality. All of these factors are potentially associated with heart disease, and are imbalanced between high and low intake groups.

Is hypertension a confounder here? Many would say no, because it is presumably in the causal pathway towards heart disease.

Collinearity: Collinearity (or multicollinearity or ill-conditioning) occurs when independent variables in a regression are so highly correlated that it becomes difficult or impossible to distinguish their individual effects on the dependent variable.

Thus, collinearity can be viewed as an extreme case of confounding, when essentially the same variable is entered into a regression equation twice, or when two variables contain exactly the same information as two other variables, and so on.

Example: It makes no sense for an independent variable to appear in a regression equation more than once. Suppose some independent variable (say, Y = height in

inches) is regressed on X_1 = age in children, and the resulting regression equation is (no claim to reality here!)

$$height = 20 + 3 \times age$$

Now suppose we attempt to fit an equation in which age appears twice as a independent variable. Suppose $X_2 = age2$ is an exact copy of the first age variable, and we attempt to fit a regression equation that includes both avriables, as follows:

$$height = alpha + \beta_1 age + \beta_2 age2$$

Note that there is no unique solution to this equation. For example, all of the above "solutions" are equivalent:

$$height = 20 + 3 \times age + 0 \times age2$$

or

$$height = 20 + 0 \times age + 3 \times age2$$

or

$$height = 20 + 1 \times age + 2 \times age2$$

or

$$height = 20 + 1.5 \times age + 1.5 \times age2$$

and so on. Note that all of these equations will produce exactly the same predictions.

The problem is that age and age2 are "collinear" variables, meaning that they each give exactly the same information. Most computer programs will either give an error message or at least a warning message if you include two collinear variables, as there is no unique choice for β_1 and β_2, an infinite number of choices being equally good. In fact, any choice in which $\beta_1 + \beta_2 = 3$ are all perfectly equivalent.

Note that collinearity does not affect the ability of a regression equation to predict the response, all of the above equations will make exactly the same predictions. However,

collinearity poses a huge problem if the objective of the study is to estimate the individual effects of each independent variable.

Strictly speaking, "collinear" implies an **exact** linear relationship between variables. For example, if age is age in years, but age 2 is age in months, they are collinear because $age2 = age \times 12$.

In general, it is not necessary to have "perfect" collinearity to cause severe problems, two variables which are highly correlated will cause a "near-collinearity" problem. This is why I suggested to always look at a correlation matrix of all variables before running a multiple linear regression, so that such potential problems can be flagged in advance.

In practice, collinearity or high correlations among independent variables will generally have the following effects:

1. Regression coefficients will change dramatically according to whether other variables are included or excluded from the model. For example, think about whether age2, a copy of the age variable, is included or excluded in a model that includes age.

2. The standard errors of the regression coefficients will tend to be large, since the beta coefficients will not be accurately estimated. In extreme cases, regression coefficients for collinear variables will be large in magnitude with signs that seem to be assigned at random. If you see "non-sensical" coefficients and SD's, collinearity should be immediately suspected as a possible cause.

3. Predictors with known, strong relationships to the response will not necessarily have their regression coefficients accurately estimated.

4. **Tolerance:** If variables are perfectly collinear, the coefficient of determination $R2$ will be 1 when any one of them is regressed upon the others. This is the motivation behind calculating a variable's "tolerance", a measure of collinearity. Each predictor can regressed on the other predictors, and its tolerance is defined as $1 - R2$. A small value of the tolerance indicates that the variable under consideration is almost a perfect linear combination of the independent variables already in the equation, and so not all these variables need to be added to the equation. Some statisticians suggest that a tolerance less than 0.1 deserves attention, although this is somewhat arbitrary.

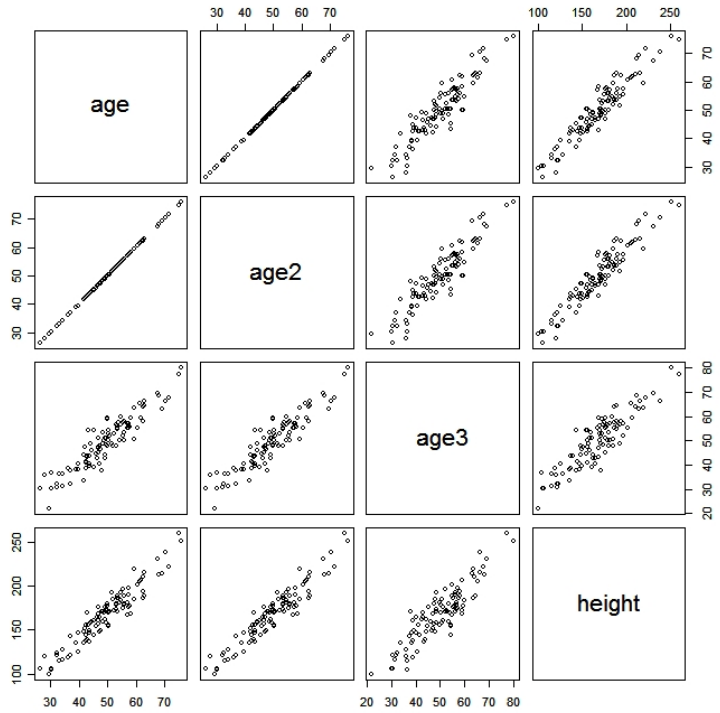
5. The tolerance is sometimes reexpressed as the Variance Inflation Factor (VIF), the inverse of the tolerance (= 1/tolerance). Tolerances of 0.10 or less become VIFs of 10 or more.

6. In general confidence intervals for regression coefficient from highly correlated variables will be wider than if the predictors were uncorrelated.

7. If the low value of tolerance is accompanied by large standard errors and thus wide confidence intervals, another study may be necessary to sort things out, unless subject matter knowledge can be used to eliminate some variables from the regression from "theory" alone.

Let's look at some examples, and see what happens in such situations:

```
#  Create an age variable with n=100

age <- round(rnorm(100, mean=50, sd=10), 2)

#  Create an exact copy of this age variable

age2 <- age

#  Create another version of this age variable,
#  but add a small amount of error, so not an exact copy

age3 <- age + rnorm(100, mean=0, sd=4)

#  Create an dependent variable that depends on age, with some error

height <- 20 + 3*age +rnorm(100, mean=0, sd=10)

#  Insert all variables into a data frame

height.dat <- data.frame(age, age2, age3, height)

#  Check correlation matrix of these four variables

cor(height.dat)

              age      age2      age3    height
age    1.0000000 1.0000000 0.9265653 0.9497912
age2   1.0000000 1.0000000 0.9265653 0.9497912
age3   0.9265653 0.9265653 1.0000000 0.8973626
height 0.9497912 0.9497912 0.8973626 1.0000000

# Look at some scatter plots

pairs(height.dat)
```



Note the extremely high correlations fund here, with *age* and *age*2 perfectly correlated (correlation = 1), and *age* and *age*3 quite highly correlated. Note also good correlations between all age variables and the outcome *height*.

Thus, we have created the perfect conditions for both collinearity and confounding (admittedly a bit artificial here) to occur.

So, let's see what happens if we try some regressions in R:

```
#  First regressions for each variable separately:

> regression.out <- lm(height ~ age)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age)

Residuals:
     Min       1Q   Median       3Q      Max
-24.0449  -8.6023   0.9563   6.7564  23.3206

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 24.78685    4.86760   5.092 1.71e-06 ***
age          2.87555    0.09569  30.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.864 on 98 degrees of freedom
Multiple R-Squared: 0.9021,     Adjusted R-squared: 0.9011
F-statistic: 903.1 on 1 and 98 DF,  p-value: < 2.2e-16

$intercept.ci
[1] 15.12726 34.44645

$slopes.ci
[1] 2.685659 3.065444

-----------------------------------------------

> regression.out <- lm(height ~ age2)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age2)

Residuals:
     Min       1Q   Median       3Q      Max
-24.0449  -8.6023   0.9563   6.7564  23.3206

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.78685    4.86760   5.092 1.71e-06 ***
age2         2.87555    0.09569  30.051  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.864 on 98 degrees of freedom
Multiple R-Squared: 0.9021,     Adjusted R-squared: 0.9011
F-statistic: 903.1 on 1 and 98 DF,  p-value: < 2.2e-16

$intercept.ci
[1] 15.12726 34.44645

$slopes.ci
[1] 2.685659 3.065444
```

```
-------------------------------------------

> regression.out <- lm(height ~ age3)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age3)

Residuals:
      Min        1Q    Median        3Q       Max
-36.17548 -10.52226  -0.04571   9.48116  34.08627

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.2819     6.5937   5.806 7.91e-08 ***
age3          2.6032     0.1293  20.130  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.91 on 98 degrees of freedom
Multiple R-Squared: 0.8053,     Adjusted R-squared: 0.8033
F-statistic: 405.2 on 1 and 98 DF,  p-value: < 2.2e-16


$intercept.ci
[1] 25.19693 51.36687

$slopes.ci
[1] 2.346546 2.859789
```

Note the identical results for *age* and *age*2, expected since these two variables are in fact identical copies of each other. Here the CIs are include the known true values for α, which was 20, and β, which was 3. Note also the imperfect estimates for age3, where the CIs for both the intercept and slope did not include the known true values. This is because of the measurement error, where age3 is not really the true age, but the age with an error added. Looking ahead, we will later see how to adjust for this measurement error. The point for this lecture, however, was imply to create another independent variable highly but not perfectly correlated with age.

Now, what happens if we try to regress two or more variables at a time, when they are either perfectly or almost perfectly correlated to each other?

```
> regression.out <- lm(height ~ age + age2)
```

```
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age + age2)

Residuals:
     Min       1Q   Median       3Q      Max
-24.0449  -8.6023   0.9563   6.7564  23.3206

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.78685    4.86760   5.092 1.71e-06 ***
age          2.87555    0.09569  30.051  < 2e-16 ***
age2              NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.864 on 98 degrees of freedom
Multiple R-Squared: 0.9021,     Adjusted R-squared: 0.9011
F-statistic: 903.1 on 1 and 98 DF,  p-value: < 2.2e-16

$intercept.ci
[1] 24.59696 24.97675

$slopes.ci
           [,1]     [,2]
[1,]  -7.229775 12.98088
[2,] -54.767366 64.50257

-----------------------------------------

> regression.out <- lm(height ~ age + age3)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age + age3)

Residuals:
     Min       1Q   Median       3Q      Max
-24.4084  -7.6603   0.2975   6.7863  21.6914

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  24.1934     4.8563   4.982 2.74e-06 ***
age           2.5321     0.2529  10.011  < 2e-16 ***
age3          0.3551     0.2423   1.465    0.146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.807 on 97 degrees of freedom
Multiple R-Squared: 0.9042,     Adjusted R-squared: 0.9022
F-statistic: 457.9 on 2 and 97 DF,  p-value: < 2.2e-16

$intercept.ci
[1] 14.55507 33.83179

$slopes.ci
           [,1]      [,2]
[1,]  2.0301561 3.0341345
[2,] -0.1258707 0.8361079

-----------------------------------------

> regression.out <- lm(height ~ age + age2 + age3)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age + age2 + age3)

Residuals:
     Min       1Q   Median       3Q      Max
-24.4084  -7.6603   0.2975   6.7863  21.6914

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.1934     4.8563   4.982 2.74e-06 ***
age           2.5321     0.2529  10.011  < 2e-16 ***
age2              NA         NA      NA       NA
age3          0.3551     0.2423   1.465    0.146
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.807 on 97 degrees of freedom
Multiple R-Squared: 0.9042,     Adjusted R-squared: 0.9022
F-statistic: 457.9 on 2 and 97 DF,  p-value: < 2.2e-16

$intercept.ci
```

```
[1] 23.69144 24.69542

$slopes.ci
           [,1]      [,2]
[1,]   2.051156  3.013135
[2,]  -9.532557 10.242794
[3,] -15.013555 24.726103
```

We note several points:

1. In the presence of perfect collinearity, R cleverly notices the collinearity, prints a message to this effect, and then automatically deletes one of the variables from the model, leaving a reasonable model where parameters are estimated. So, a model with age and age2 provides the same results as a model with just age, because age2 is automatically dropped, and a model with age, age2 and age3 is the same as a model with just age and age3, again because age2 is dropped.
2. Not all programs do this, some will just return an error message and stop.
3. Note that our customized R function did not work properly. This is because the order we are expecting for outputs in the object "regression.out" has changed because of the dropped variable. We must be careful about such special cases.
4. When age and age3 are included, note that the coefficient of age is biased downwards, due to presence of similar variable age3. We are lucky in that the CI for age still includes 3, but this will not always be the case. For example, create a stronger age3 variable, and observe the difference:

```
#  Before, we used age3 <- age + rnorm(100, mean=0, sd=4)
#  now try with smaller SD for age3:

> age3 <- age + rnorm(100, mean=0, sd=1)
>
> regression.out <- lm(height ~ age + age3)
>
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = height ~ age + age3)

Residuals:
    Min      1Q  Median      3Q     Max
-26.808  -8.351   2.029   7.041  22.338
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.0917     4.8571   4.960    3e-06 ***
age           1.4109     0.9690   1.456    0.149
age3          1.4757     0.9716   1.519    0.132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.799 on 97 degrees of freedom
Multiple R-Squared: 0.9044,     Adjusted R-squared: 0.9024
F-statistic: 458.7 on 2 and 97 DF,  p-value: < 2.2e-16

$intercept.ci
[1] 14.45173 33.73166

$slopes.ci
           [,1]     [,2]
[1,] -0.5124352 3.334136
[2,] -0.4526711 3.403996
```

Notice now that the effect of adding age3 to the model with age is much more extreme, the effect of 3 essentially becomes split between the two very similar variables, roughly 1.5 each. Note also the much larger SEs for age, for example, changed from 0.25 to almost 1, a four-fold increase in uncertainty. Note that this caused very wide CIs for both age and age3.

Example

Here is another example, using real data:

In 1993, an experiment was conducted at Ohio State University exploring the relationship between heart rate and the frequency at which that person stepped up and down on steps of various heights. The response variable, heart rate, was measured in beats per minute. There were two different step heights: 5.75 inches (coded as 0), and 11.5 inches (coded as 1). There were three rates of stepping: 14 steps/min. (coded as 0), 21 steps/min. (coded as 1), and 28 steps/min. (coded as 2). This resulted in six possible height/frequency combinations. Each subject performed the activity for three minutes. Subjects were kept on pace by the beat of an electric metronome. One experimenter counted the subject's pulse for 20 seconds before and after each trial. The subject always rested between trials until her or his heart rate returned to close to the beginning rate. Another experimenter kept track of the time spent stepping.

Each subject was always measured and timed by the same pair of experimenters to reduce variability in the experiment. Each pair of experimenters was treated as a block.

Order:	the overall performance order of the trial
Block:	the subject and experimenters' block number
Height:	0 if step at the low (5.75") height, 1 if at the high (11.5") height
Frequency:	the rate of stepping. 0 if slow (14 steps/min), 1 if medium (21 steps/min), 2 if high (28 steps/min)
RestHR:	the resting heart rate of the subject before a trial, in beats per minute
HR:	the final heart rate of the subject after a trial, in beats per minute

Block	Height	Frequency	RestHR	HR
2	0	0	60	75
2	0	1	63	84
2	1	2	69	135
2	1	0	69	108
2	0	2	69	93
4	1	1	96	141
4	1	0	87	120
4	0	0	90	99
4	1	2	93	153
4	0	2	87	129
3	1	1	72	99
3	0	1	69	93
3	1	0	78	93
3	0	2	72	99
3	1	2	78	129
5	0	0	87	93
1	1	1	87	111
6	1	2	81	120
5	0	2	75	123
1	0	1	81	96
6	1	0	84	99
1	1	0	84	99
5	1	1	90	129
6	0	1	75	90
1	0	0	78	87
6	0	0	84	84
5	0	1	90	108
1	0	2	78	96
6	1	1	84	90
5	1	2	90	147

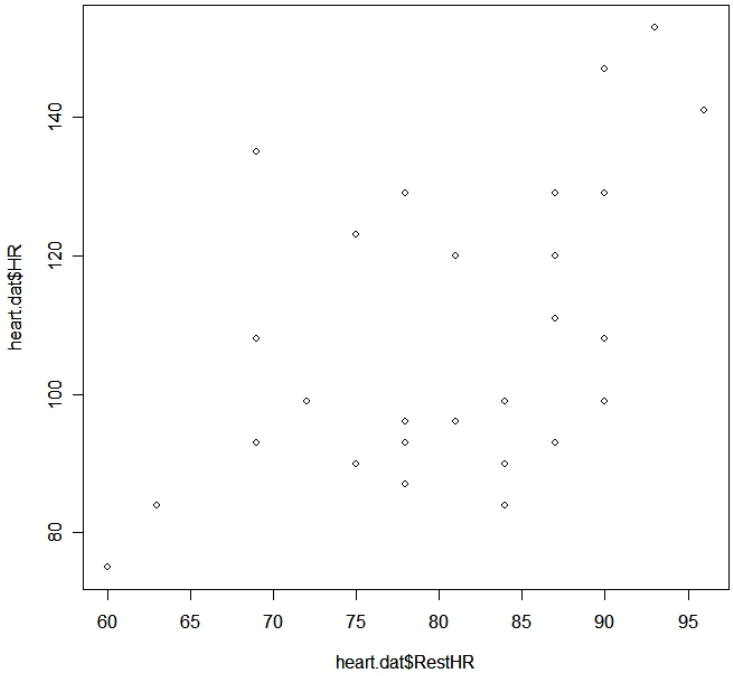
We will now analyze these data in R. We will follow these steps:

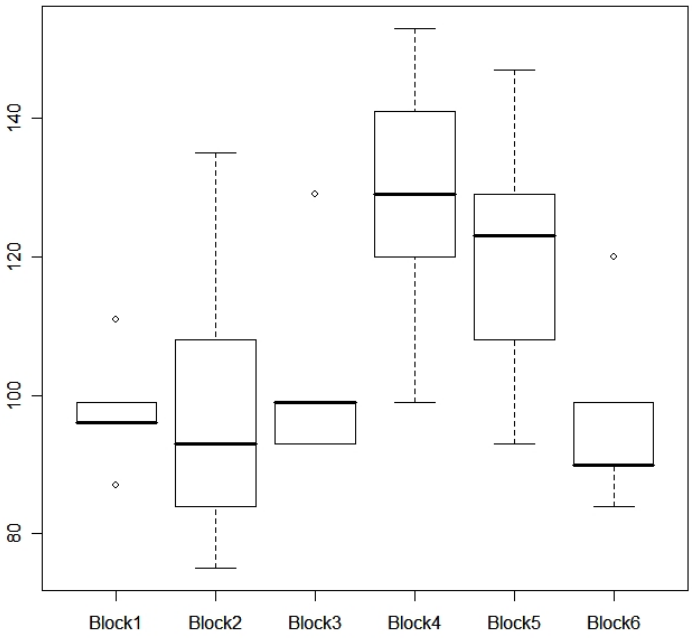
1. Enter the data.
2. Look at some simple descriptive statistics
3. Look at univariate regressions for each variable.
4. Compare these coefficients to those from a full multivariate model.
5. Possibly eliminate variables from the model that look unimportant.
6. State conclusions by looking at confidence intervals, including stating whether there may be any confounding or not.

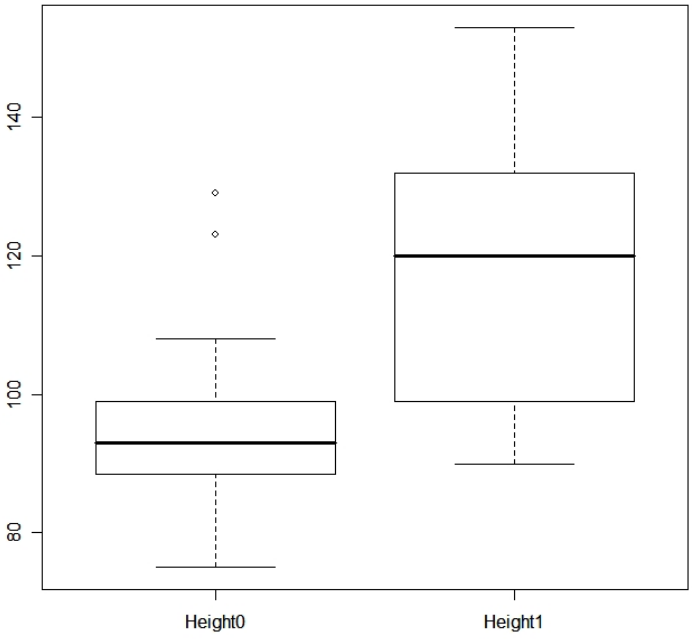
```
#  Enter the data which is saved somewhere as hr.txt

> heart.dat <- read.table(file="g:\\hr.txt", header=T)

# Take a quick look at the data, which seems fine.

> heart.dat
   Block Height Frequency RestHR  HR
1      2      0         0     60  75
2      2      0         1     63  84
3      2      1         2     69 135
4      2      1         0     69 108
5      2      0         2     69  93
6      4      1         1     96 141
7      4      1         0     87 120
8      4      0         0     90  99
9      4      1         2     93 153
10     4      0         2     87 129
11     3      1         1     72  99
12     3      0         1     69  93
13     3      1         0     78  93
14     3      0         2     72  99
15     3      1         2     78 129
16     5      0         0     87  93
17     1      1         1     87 111
18     6      1         2     81 120
19     5      0         2     75 123
20     1      0         1     81  96
21     6      1         0     84  99
22     1      1         0     84  99
```

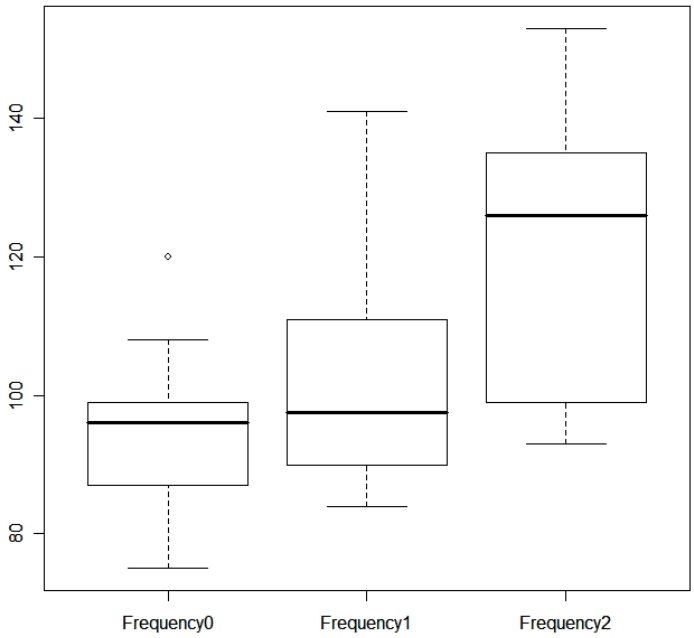
```
23     5      1         1     90 129
24     6      0         1     75  90
25     1      0         0     78  87
26     6      0         0     84  84
27     5      0         1     90 108
28     1      0         2     78  96
29     6      1         1     84  90
30     5      1         2     90 147

> summary(heart.dat)
     Block         Height      Frequency     RestHR            HR
 Min.   :1.0   Min.   :0.0   Min.   :0   Min.   :60.00   Min.   : 75.0
 1st Qu.:2.0   1st Qu.:0.0   1st Qu.:0   1st Qu.:72.75   1st Qu.: 93.0
 Median :3.5   Median :0.5   Median :1   Median :81.00   Median : 99.0
 Mean   :3.5   Mean   :0.5   Mean   :1   Mean   :80.00   Mean   :107.4
 3rd Qu.:5.0   3rd Qu.:1.0   3rd Qu.:2   3rd Qu.:87.00   3rd Qu.:122.3
 Max.   :6.0   Max.   :1.0   Max.   :2   Max.   :96.00   Max.   :153.0

#  Notice that some factor variables are not being treated that way, so:
#  Recall case is important in R, frequency is NOT the same as Frequency!

> heart.dat$Block <- as.factor(heart.dat$Block)
> heart.dat$Height <- as.factor(heart.dat$Height)
> heart.dat$Frequency <- as.factor(heart.dat$Frequency)

# Redo summary, notice that it is now corrected.

> summary(heart.dat)
 Block Height Frequency     RestHR            HR
 1:5   0:15   0:10      Min.   :60.00   Min.   : 75.0
 2:5   1:15   1:10      1st Qu.:72.75   1st Qu.: 93.0
 3:5          2:10      Median :81.00   Median : 99.0
 4:5                    Mean   :80.00   Mean   :107.4
 5:5                    3rd Qu.:87.00   3rd Qu.:122.3
 6:5                    Max.   :96.00   Max.   :153.0

#  Note balance in the design, good for defeating confounding
#  because all variables are balanced.

 # Scatter plot for continuous variables, boxplots for the rest:

 > plot(heart.dat$RestHR, heart.dat$HR)

#  To simplify typing
```

```
> attach(heart.dat)

> boxplot(list(Block1=HR[Block==1], Block2=HR[Block==2],
  Block3=HR[Block==3], Block4=HR[Block==4],
  Block5=HR[Block==5], Block6=HR[Block==6]))

> boxplot(list(Height0=HR[Height==0], Height1=HR[Height==1]))

> boxplot(list(Frequency0=HR[Frequency==0], Frequency1=HR[Frequency==1],
  Frequency2=HR[Frequency==2]))
```




140
120
100
80
Block1
Block2
Block3
Block4
Block5
Block6


140
120
100
80
Height0
Height1




```
# Univariate regression for each variable
# [For brevity, just show a portion of the results here]

> regression.out <- lm(HR ~ RestHR, data=heart.dat)
> multiple.regression.with.ci(regression.out)
$regression.table
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.6861    28.8532   0.509  0.61475
RestHR        1.1589     0.3584   3.234  0.00313 **

-------------------------------------------

> regression.out <- lm(HR ~ Block, data=heart.dat)
> multiple.regression.with.ci(regression.out)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   97.800      7.964  12.281 7.73e-12 ***
Block2         1.200     11.262   0.107   0.9160
Block3         4.800     11.262   0.426   0.6738
Block4        30.600     11.262   2.717   0.0120 *
Block5        22.200     11.262   1.971   0.0603 .
Block6        -1.200     11.262  -0.107   0.9160
```

```
#  Note that category 1 is the "reference category",
#  whose mean is given by the intercept.

#  Also note that only category 4 reaches "significance",
#  very common for this to happen for categorical variables.

-------------------------------------------

> regression.out <- lm(HR ~ Height, data=heart.dat)
> multiple.regression.with.ci(regression.out)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   96.600      4.531  21.321   <2e-16 ***
Height1       21.600      6.408   3.371   0.0022 **

# Category 0 is the reference.

-------------------------------------------

> regression.out <- lm(HR ~ Frequency, data=heart.dat)
> multiple.regression.with.ci(regression.out)
$regression.table

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   95.700      5.576  17.164 4.72e-16 ***
Frequency1     8.400      7.885   1.065  0.29617
Frequency2    26.700      7.885   3.386  0.00219 **

# Category 0 is the reference, only Frequency2
# is "significant".
-------------------------------------------

#  Now compare these estimated coefficients with
#  the full model, to check for any differences in
#  estimated parameters

> regression.out <- lm(HR ~ RestHR + Block + Height + Frequency, data=heart.dat)
> multiple.regression.with.ci(regression.out)
$regression.table

Call:
lm(formula = HR ~ RestHR + Block + Height + Frequency, data = heart.dat)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-15.8984  -4.0292   0.2189   5.9545   9.7953

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  65.5002    34.8943   1.877  0.07517 .
RestHR        0.1874     0.4353   0.431  0.67137
Block2        0.9662     8.2908   0.117  0.90838
Block3       -2.8990     6.1834  -0.469  0.64426
Block4       21.5927     6.0529   3.567  0.00193 **
Block5       16.3019     5.3150   3.067  0.00608 **
Block6       -5.3626     4.8664  -1.102  0.28356
Height1      20.8128     3.5752   5.821 1.07e-05 ***
Frequency1    9.2031     3.4756   2.648  0.01544 *
Frequency2   24.9921     3.4845   7.172 6.03e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 7.611 on 20 degrees of freedom
Multiple R-Squared: 0.9044,     Adjusted R-squared: 0.8614
F-statistic: 21.03 on 9 and 20 DF,  p-value: 2.533e-08


$intercept.ci
[1]  -7.288011 138.288404

$slopes.ci
             [,1]      [,2]
 [1,]  -0.720578  1.095455
 [2,] -16.328035 18.260518
 [3,] -15.797328  9.999389
 [4,]   8.966474 34.218904
 [5,]   5.215017 27.388712
 [6,] -15.513617  4.788494
 [7,]  13.354991 28.270622
 [8,]   1.953209 16.453072
 [9,]  17.723583 32.260715
```

For ease of comparison, let create the following table:

	Multivariate		Univariate	
Variable	Estimate	Std. Error	Estimate	Std. Error
(Intercept)	65.5002	34.8943		
RestHR	0.1874	0.4353	1.1589	0.3584
Block2	0.9662	8.2908	1.200	11.262
Block3	-2.8990	6.1834	4.800	11.262
Block4	21.5927	6.0529	30.600	11.262
Block5	16.3019	5.3150	22.200	11.262
Block6	-5.3626	4.8664	-1.200	11.262
Height1	20.8128	3.5752	21.600	6.408
Frequency1	9.2031	3.4756	8.400	7.885
Frequency2	24.9921	3.4845	26.700	7.885

From the above table, we note the following:

1. There is only mild confounding for most variables, but RestHR changed very substantially. While most estimates do move around a bit, for the most part movement is all easily within the confidence intervals, and many SEs were large anyway. There was substantial confounding with RestHR, however. Looking back at the data set, one suspects that subjects did not fully return to their true Resting HEart Rates after each trial. For example, looking at the data from Block2, we see that RestHR changes from 60 to 69 as exercise gets more difficult. Thus, after accounting for the other variables, RestHR appears to become less important a predictor.

2. The SEs are much smaller from the multivariate model compared to the univariate model. This often happens: if "good" independent variables are added to a model, then the residual SD decreases, so SEs, which depend in large part on the residual SD, decrease.

3. From confidence intervals, we see that Block4 is at least 8 beats, and up to 34 beats faster than Block1 (reference), which seems clinically important. Height makes at least a 13 beat difference, and both Frequency1 and Frequency2 are different from Frequency0, although not all values in the CI's are clinically relevant for Frequency1.

4. **Important:** Other categories are **inconclusive**, due to wide CIs, so more research is needed for them. We **cannot** conclude that these variables are unimportant!! This is an **extremely common misconception** made by researchers everywhere!!

One way to View Confounding

Suppose we have a model with Y as the outcome, X as a dichotomous variable of main interest, and C as a potential confounder. If we ignore the confounder, then we would run this model:

$$Y = \alpha + \beta_1 X$$

From that model, the effect of a one unit change in X is simply β_1.

However, if we add the potential confounder to the model, the model becomes:

$$Y = \alpha^* + \beta_1^* X + \beta_2 C$$

where it can be shown that

$$\beta_1^* = \beta_1 - \beta_2(\overline{C}_{X=1} - \overline{C}_{X=0})$$

So, there are several possibilities:

1. If C is **not** related to X, then $\overline{C}_{X=1} - \overline{C}_{X=0} \approx 0$, and $\beta_1^* \approx \beta_1$.
2. If C is **not** related to Y, then $\beta_2 \approx 0$, and $\beta_1^* \approx \beta_1$.

   In both of the above cases, there is in fact no confounding, so, not surprisingly, β_1 stays the same, regardless as to whether C is in the model or not.

3. If C **is** in fact related to both X and Y, we have confounding, and β_1 is a biased estimate of the true effect of X on Y, owing to the confounding from C. Thus, we must include C in the model, use β_1^* as the estimate the effect of X on Y.

   The magnitude of the bias depends both on the strength of the relationship between C on Y, as measured by β_2, and on the strength of the relationship between X and C, as measured by $\overline{C}_{X=1} - \overline{C}_{X=0}$.