MA50259: Statistical Design of Investigations

Lab sheet 9: Confounding in observatioonal studies

In this practical you will learn about how to adjust for confounding in an observational study.

Confounding

Consider the following data on an observational study of 500 singleton births in a London Hospital. The data has the following recorded variables

- **id:** Identity number for mother and baby
- **bweight:** Birth weight of baby in grams
- **lowbw:** Indicator for birth weight less than 2500 grms
- **gestwks:** Gestation period in weeks
- **preterm:** Indicator for gestation period less than 37 weeks
- **matage:** Maternal age in years
- **hyp:** Indicator for maternal hypertension
- **sex:** Sex of baby: 1:Male, 2:Female

Load the data using the following command:

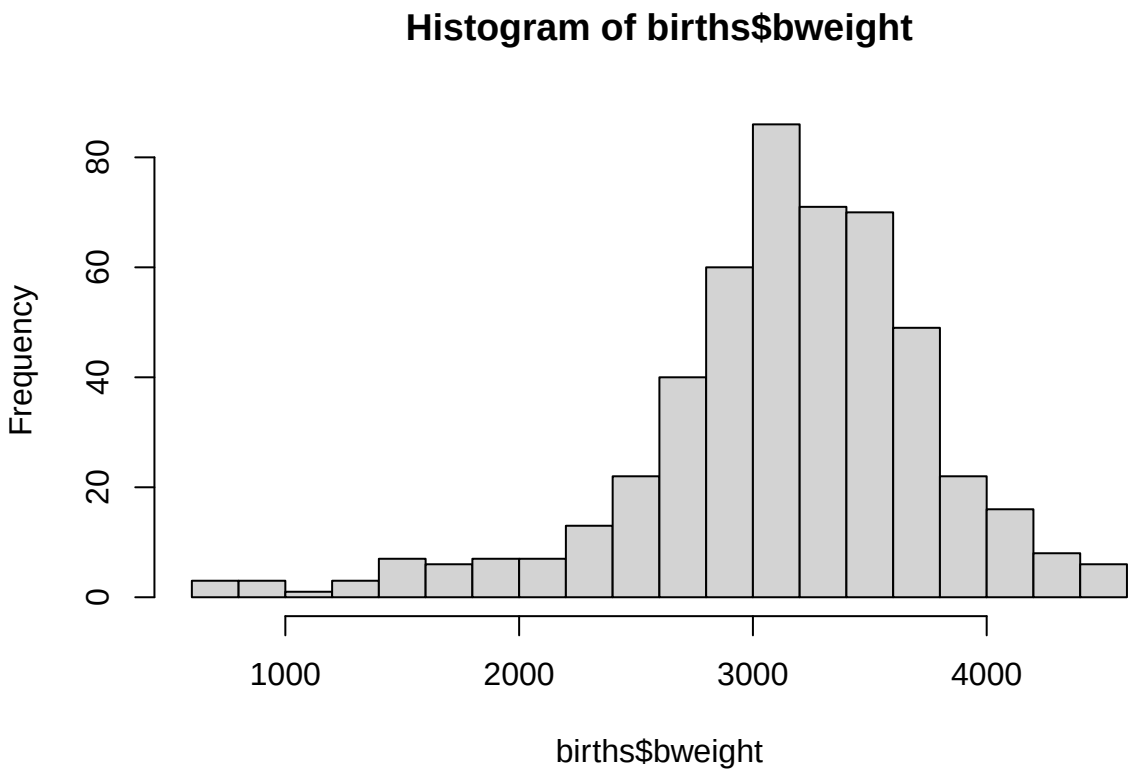
```
library(tidyverse)
births<-"http://people.bath.ac.uk/kai21/MA50259/Data/births.txt" %>% read.table(header=TRUE)
head(births)
```

```
##   id bweight lowbw gestwks preterm matage    hyp sex
## 1  1    2974     0   38.52       0     34 normal   F
## 2  2    3270     0      NA      NA     30 normal   M
## 3  3    2620     0   38.15       0     35 normal   F
## 4  4    3751     0   39.80       0     31 normal   M
## 5  5    3200     0   38.89       0     33  hyper   M
## 6  6    3673     0   40.97       0     33 normal   F
```

The outcome of interest in this study is the birt weight of the baby in grams.

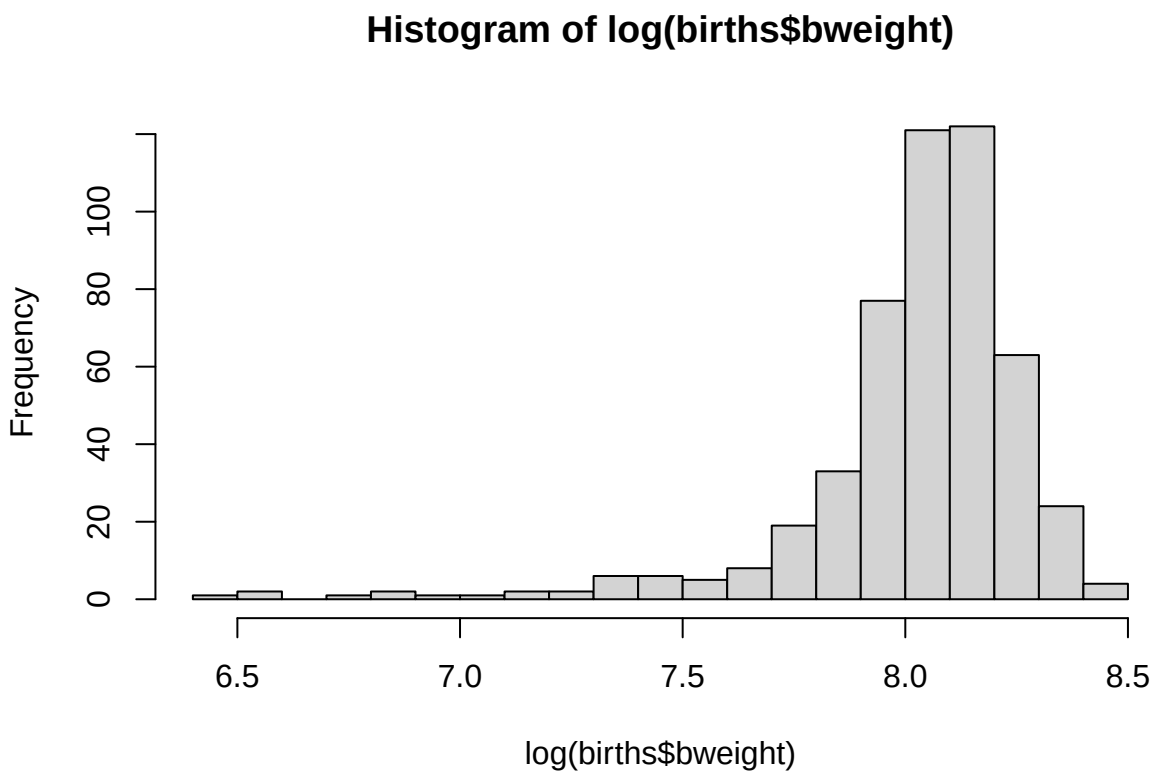
1. We will assume that the birth weight of the baby is normally distributed. Is this assumption tenable?

```
hist(births$bweight,20)
```



Solution: The histogram of the birth weight observations shows the distribution might be a bit skewed to one side compared to the normaldistribution that is symmetric. The assumption is then not tenable given the relatively large sample size. Applying a logarithmic transformation does not solve the problem

```
hist(log(births$bweight),20)
```



We can proceed with the normal assumption but be a bit cautious on the conclusions we draw, specially with respect to variance size.

2. Consider the case where the exposure (treatment) is the presence or absence of maternal hypertension. This clearly cannot be randomised, so we just observe the corresponding status. What is the average increase (or decrease) of birth weight of hypertense mothers with respect to normal mothers?

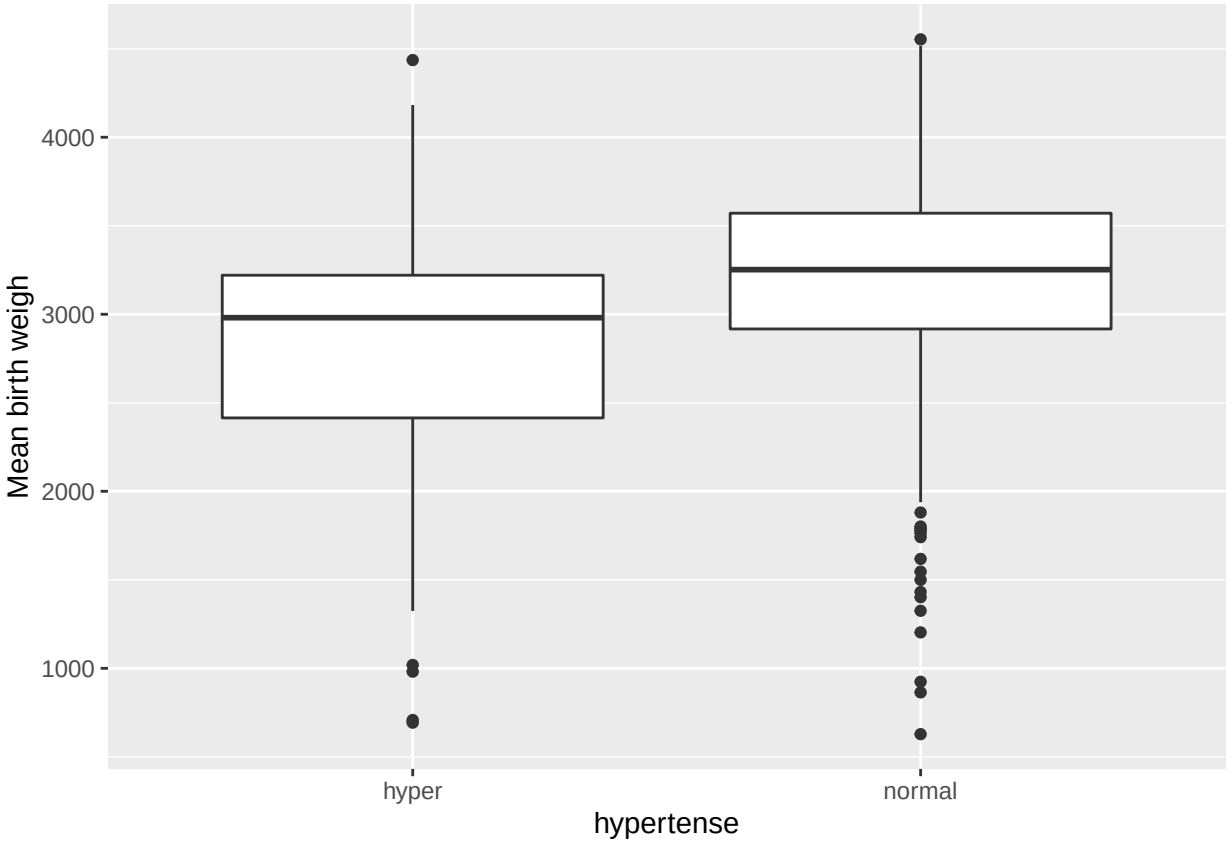
Solution:

```
avg_bweight_normal<-mean(births$bweight[births$hyp=="normal"])
avg_bweight_hyper<-mean(births$bweight[births$hyp=="hyper"])
avg_bweight_normal-avg_bweight_hyper
```

```
## [1] 430.6959
```

The average incvrease is of 430.6958723 grams and looking at the box plots below is not clear if this difference is significant as there does not seem to be enough overlap.

```
ggplot(births,aes(hyp,bweight)) + geom_boxplot() + xlab("hypertense") + ylab("Mean birth weigh")
```




```
mod0<-lm(bweight~hyp,births)
summary(mod0)
```

```
##
## Call:
## lm(formula = bweight ~ hyp, data = births)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2570.9  -286.4    69.1   383.9  1667.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2768.21      73.05  37.895  < 2e-16 ***
## hypnormal     430.70      78.95   5.455 7.73e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 619.8 on 498 degrees of freedom
## Multiple R-squared:  0.05638,    Adjusted R-squared:  0.05449
## F-statistic: 29.76 on 1 and 498 DF,  p-value: 7.729e-08
```

```
coefficients(mod0)
```

```
## (Intercept)   hypnormal
```

```
##   2768.2083    430.6959
```

3. What are the potential outcomes and how would you define the individual-level causal effect in this case?

Solution the potential outcomes are Y^{hyper} the birthweight of the baby if the mother hypertense and Y^{normal} the birthweight of the baby if the mother has normal blood pressure. The individual level causal effect can be defined in many ways but the usual way is the difference

$$Y^{normal} - Y^{hyper}$$

4. Test the hypothesis (using $\alpha = 0.05$) that there is no average effect of maternal hypertension on the birthweight of the baby

```
summary(aov(mod0))
```

```
##              Df    Sum Sq  Mean Sq F value   Pr(>F)
## hyp           1  11432670 11432670   29.76 7.73e-08 ***
## Residuals   498 191333183   384203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# alternatively
table<-summary(mod0)$coeff
table
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 2768.2083   73.04899 37.895230 7.892113e-149
## hypnormal    430.6959   78.95458  5.454983  7.728888e-08
```

```
table["hypnormal","Pr(>|t|)"]
```

```
## [1] 7.728888e-08
```

Solution: This indicates there is a significant difference.

5. Under which assumption we can consider the above average effect of maternal hypertension on the birthweight of the baby, to be causal? Is this assumption tenable?

Solution: Under the assumption that those mothers who are hypertense are exchangeable with those who have normal blood pressure. Here exchangeability is with respect to any other aspect apart from blood pressure which seems highly untenable. To understand better this last statement you should imagine the following strange assumptions are tryue:

- Normal (blood pressure) mothers have low birthweight babies as often as hypertense mothers had they not been hypertense!
- Hypertense mothers have low birthweight babies as often as normal mothers had they been hypertense!

Please note how we are dealing with hypertension as an exposure (or treatment) in the same way as we would if the exposure were a pill for example. In such hypothetical situation, we would be able to randomize hypertension to expecting mothers and in that case the above assumptions would have been true! This is of course, not possible in reality but only mathematically!

If the above assumptions are true then we can assume the effect found above is causal. This is because we can prove that the distribution of the potential outcome Y^{hyper} is the same as the conditional distribution of the actual (observed) outcome Y given hypertension, that is $Y|hyper$. The potential outcome Y^{hyper}, like one's genetic make-up, can be thought of as a fixed characteristic of a mother existing before the hypertension treatment is assigned. We can think that Y^{hyper} encodes what would have been a mother's outcome if assigned to have hypertension and thus does not depend on later being assigned to hypertension or not!

Now back to reality! the above assumptions are untenable and the effect found above cannot be seen as causual but only as associational. The best we can conclude is that there seems to be an association between hypertesion and low birth weight!

6. Is the assumption that the average effect of maternal hypertension on the birthweight of the baby within the strata defined by the sex, is causal, a tenable assumption? Are the baby gender groups comparable?

```
levels(births$sex)
```

```
## NULL
```

```
subset.male<-births$sex=="M"
subset.female<-births$sex=="F"
mod.male<-lm(bweight~hyp,births,subset=subset.male)
mod.female<-lm(bweight~hyp,births,subset=subset.female)
summary(mod.male)
```

```
##
## Call:
## lm(formula = bweight ~ hyp, data = births, subset = subset.male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2121.40  -336.00    74.75   364.02  1621.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2814.40      92.66  30.374  < 2e-16 ***
## hypnormal     496.35     101.27   4.901 1.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 607.6 on 262 degrees of freedom
## Multiple R-squared:  0.08398,    Adjusted R-squared:  0.08049
## F-statistic: 24.02 on 1 and 262 DF,  p-value: 1.672e-06
```

```
summary(mod.female)
```

```
##
```

```
## Call:
## lm(formula = bweight ~ hyp, data = births, subset = subset.female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2451.5  -235.8    79.0   383.6  1220.5
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2699.7      114.3  23.619  < 2e-16 ***
## hypnormal      379.8      122.0   3.112  0.00209 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 615.5 on 234 degrees of freedom
## Multiple R-squared:  0.03973,    Adjusted R-squared:  0.03563
## F-statistic: 9.682 on 1 and 234 DF,  p-value: 0.002092
```

```
coefficients(mod.male)
```

```
## (Intercept)   hypnormal
##   2814.3953    496.3513
```

```
coefficients(mod.female)
```

```
## (Intercept)   hypnormal
##   2699.7241    379.7734
```

```
coefficients(mod0)
```

```
## (Intercept)   hypnormal
##   2768.2083    430.6959
```

Solution: The effect of hypertension on birthweight seems to be different for male babies compared to female babies (496.35 grams compared to 379.8 grams of increase in birthweight for normal mothers). Without any expert clinical knowledge it seems that ther is no reason to assume that the assumptions above, but now within mothers with babies of the same gender, are more tenable than before. The groups are not comparable as they might differ in terms of other characteristics which were not measure (such as lifestyle of the mother).

7. What is the average effect of maternal hypertension on the birth weight of the baby after adjusting for the sex of the baby. When comparing with average effect found in question 2, what can we conclude?

```
mod.adj<-lm(bweight~hyp+sex,births)
summary(mod.adj)
```

```
##
## Call:
## lm(formula = bweight ~ hyp + sex, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -2459.89  -294.38    83.65   364.18  1581.20
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2639.81      79.11  33.368  < 2e-16 ***
## hypnormal     448.08      77.97   5.747 1.58e-08 ***
## sexM          214.99      54.83   3.921 0.000101 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 611.1 on 497 degrees of freedom
## Multiple R-squared:  0.0847, Adjusted R-squared:  0.08101
## F-statistic: 22.99 on 2 and 497 DF,  p-value: 2.811e-10
```

```
coefficients(mod.adj)
```

```
## (Intercept)   hypnormal        sexM
##   2639.8097    448.0817    214.9931
```

Solution:

After adjusting for gender of the baby, the average effect of maternal hypertension on the birth weight of the baby is an increase of 448.08 grams for normal mothers with respect to hypertense mother. Note this value is somewhere in between the stratum specific estimates above of 496.35 grams and 379.8 grams and also differs from only slightly with the effect obtained in question 2 of 430.7. The association between hypertension and birtweight seems not be distorted by the gender of the baby born, in other words it does not seem to be a confounder.

From the definition of confounder we can call the variable: gender of the baby, a confounder only if it is tenable to assume that the gender of the baby is associated with the birthweight.

A confounder in this case can clearly be the age of the mother. For simplicity we dichotomise this variable by cutting in the median value of 34 years of age so that we have

```
births$over.median.age<-rep(0,length(births$bweight))
births$over.median.age<-as.numeric(births$matage>=34)
subset.over<-births$over.median.age==1
subset.below<-births$over.median.age==0
mod.over<-lm(bweight~hyp,births,subset=subset.over)
mod.below<-lm(bweight~hyp,births,subset=subset.below)
mod.all<-lm(bweight~hyp+over.median.age,births)
mod.all2<-lm(bweight~hyp+matage,births)
coefficients(mod.over)
```

```
## (Intercept)   hypnormal
##   2869.5588    303.5532
```

```
coefficients(mod.below)
```

```
## (Intercept)   hypnormal
##   2677.5263    557.6029
```

```
coefficients(mod.all)
```

```
##     (Intercept)       hypnormal over.median.age
##      2779.83659       433.45111       -24.62453
```

```
coefficients(mod.all2)
```

```
##   (Intercept)     hypnormal        matage
## 2770.24818829  430.73946765   -0.06104304
```

```
coefficients(mod0)
```

```
## (Intercept)   hypnormal
##   2768.2083    430.6959
```

In any case the average effect seems almost identical to the unadjusted effect obtained in question 2.However one might stil argue that maternal age is a confounder since is clearly associated with the outcome (birthweight of the baby) and with the the exposure (hypertension). This is the best we can as there is reallyno statistical test for confounding.

8. Consider all the previous questions again but now with the exposure given by the gestation in weeks, which clearly also cannot be randomised!

```
mod0<-lm(bweight~gestwks,births)
summary(mod0)
```

```
##
## Call:
## lm(formula = bweight ~ gestwks, data = births)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1698.40  -280.14    -3.64   287.61  1382.24
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4489.140    340.899  -13.17   <2e-16 ***
## gestwks       196.973      8.788   22.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 449.7 on 488 degrees of freedom
##   (10 observations deleted due to missingness)
## Multiple R-squared:  0.5073, Adjusted R-squared:  0.5062
## F-statistic: 502.4 on 1 and 488 DF,  p-value: < 2.2e-16
```

```
coefficients(mod0)
```

```
## (Intercept)     gestwks
##  -4489.1398    196.9726
```

Solution: From the output above there seems to be a positive association in the sense that the more weeks of gestation then the more weight on the baby (to some extent). We can also read the association in the other direction where the less weeks of gestation then the lower the birth weight! Again this association would in fact be causal if exchageability holds in the sense described in the previous questions. For example assumption would hold in the case of exchangeoability (aka ignorability):

- Mothers who deliver babies early have low birthweight babies as often as mothers who deliver babies late had they not been late in delivering their babies!

Again this assumption seem untenable and the effect seen can only be interpreted as an association.

Now, after adjusting for maternal age it seems the association does not change but we can still keep this variable as a confounder !

```
mod1<-lm(bweight~gestwks+matage,births)
coefficients(mod1)
```

```
##   (Intercept)       gestwks        matage
## -4.490385e+03  1.969710e+02  3.838254e-02
```

```
coefficients(mod0)
```

```
## (Intercept)     gestwks
##  -4489.1398    196.9726
```