

Running a Simulation Study

An Example of a statistical simulation

Small scale simulation study to investigate impact of measurement error on (continuous) exposure and/or (continuous) confounding variable. In epidemiological studies of the relation between an exposure and an outcome, this relation is often estimated using regression analysis. As an example, we consider a study of the association between glycated haemoglobin (HbA1c) levels and systolic blood pressure assessed using linear regression. Data from 5092 subjects in the 2015–2016 National Health and Nutrition Examination Survey (NHANES) are used to obtain an estimate of the effect of HbA1c on systolic blood pressure, while adjusting for age, gender and body mass index (BMI).

Analysis

After adjustment for age and gender, it was estimated that HbA1c increases systolic blood pressure by 1.13 mmHg (95% CI 0.73 to 1.52) per unit increase in HbA1c.

```
summary(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
```

Call:

```
lm(formula = bp ~ HbA1C + age + as.factor(sex), data = dc)
```

Residuals:

Systolic: Blood pres (1st rdg) mm Hg

Min	1Q	Median	3Q	Max
-49.887	-10.509	-1.378	8.491	107.583

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	98.75149	1.21418	81.332	< 2e-16 ***

```

HbA1C          1.12638    0.20291    5.551 2.98e-08 ***
age            0.44486    0.01284   34.648 < 2e-16 ***
as.factor(sex)2 -3.24792    0.45164   -7.191 7.34e-13 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.1 on 5088 degrees of freedom

Multiple R-squared: 0.2305, Adjusted R-squared: 0.23

F-statistic: 508 on 3 and 5088 DF, p-value: < 2.2e-16

```
confint(lm(bp ~ HbA1C + age + as.factor(sex), data=dc))
```

```

                2.5 %      97.5 %
(Intercept)    96.3711755 101.1317982
HbA1C           0.7285836  1.5241825
age            0.4196932  0.4700355
as.factor(sex)2 -4.1333281 -2.3625106

```

Additional adjustment for BMI resulted in a considerable change in the effect estimate: HbA1c was estimated to increase blood pressure by 0.75 mmHg (95% CI 0.35 to 1.16) per unit increase in HbA1c.

```
summary(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
```

Call:

```
lm(formula = bp ~ HbA1C + bmi + age + as.factor(sex), data = dc)
```

Residuals:

Systolic: Blood pres (1st rdg) mm Hg

```

      Min       1Q   Median       3Q      Max
-51.068 -10.251  -1.504    8.264  107.410

```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   92.65583    1.39320   66.506 < 2e-16 ***
HbA1C          0.75177    0.20596    3.650 0.000265 ***
bmi            0.28632    0.03282    8.724 < 2e-16 ***
age            0.44586    0.01275   34.979 < 2e-16 ***
as.factor(sex)2 -3.63115    0.45049   -8.060 9.4e-16 ***

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.98 on 5087 degrees of freedom

Multiple R-squared: 0.2418, Adjusted R-squared: 0.2412

F-statistic: 405.7 on 4 and 5087 DF, p-value: < 2.2e-16

```
confint(lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc))
```

	2.5 %	97.5 %
(Intercept)	89.9245592	95.3871089
HbA1C	0.3479966	1.1555348
bmi	0.2219815	0.3506673
age	0.4208695	0.4708464
as.factor(sex)2	-4.5143014	-2.7479929

Run the simulation study

One way to investigate the possible impact of measurement error is through a small simulation study. For the purpose of this example, the original recordings in the NHANES data were assumed to be measured without error. Then, in addition, new artificial variables were created that represented HbA1c and BMI, but for the situation in which these are measured with error. To create these variables, measurement error was artificially added to the exposure variable (HbA1c) and/or the confounding variable (BMI). These errors were drawn from a normal distribution with a mean zero and were independent of all variables considered. The variance of the normal distribution, defining the amount of measurement error added, was altered for different scenarios. Scenarios ranged from no measurement error on either HbA1c or BMI (reference scenario) to 50% of the variance in HbA1c and/or BMI attributable to measurement error. To minimise the impact of simulation error, each scenario was repeated 1000 times and the results were averaged per scenario over these 1000 repetitions.

```
ref <- lm(bp ~ HbA1C + bmi + age + as.factor(sex), data=dc)$coef[2]
n.sim <- 1e3
pb <- txtProgressBar(min=1, max= 1e3, style=3, 50)
perc.me.exp <- seq(0,.5,.1)
perc.me.conf<- seq(0,.5,.1)
scenarios <- expand.grid(perc.me.exp,perc.me.conf)
var.exp <- var(dc$HbA1C)
var.conf <- var(dc$bmi)
n <- dim(dc)[1]
```

```

beta.hat <- matrix(ncol=dim(scenarios)[1], nrow=n.sim)

for (k in 1:n.sim){
  set.seed(k)
  for (i in 1:dim(scenarios)[1]){
    var.me.exp <- var.exp*scenarios[i,1]/(1-scenarios[i,1])
    var.me.conf <- var.conf*scenarios[i,2]/(1-scenarios[i,2])
    dc$HbA1C.me <- dc$HbA1C + rnorm(dim(dc)[1], 0, sqrt(var.me.exp))
    dc$bmi.me <- dc$bmi + rnorm(dim(dc)[1], 0, sqrt(var.me.conf))
    beta.hat[k,i] <- lm(bp ~ HbA1C.me + age + bmi.me + as.factor(sex), data=dc)$coef[2]
    setTxtProgressBar(pb, k)
  }}

close(pb)

```

Results

Thus figure shows the impact of measurement error on HbA1c and/or BMI on the estimate of the regression coefficient of HbA1c. The relation between HbA1c and systolic blood pressure was attenuated when measurement error was added to HbA1c, but not when measurement error was added to BMI. The association became stronger as measurement error was added solely to the confounding variable BMI. The reason for this effect is that, with increasing levels of measurement error on BMI, adjustment for the confounding due to BMI becomes less efficient and the effect estimate gets closer to the unadjusted estimate (1.13 mmHg). Due to measurement error, a type of residual confounding is introduced. In the case of measurement error on HbA1c as well as BMI, both phenomena play a role and may cancel each other out. In this study, measurement error on HbA1c seemed more influential than measurement error on BMI.

```

tot.mat <- cbind(100*scenarios,apply(beta.hat,2,mean))
colnames(tot.mat) <- c("me.exp","me.conf","estimate")
ggplot(tot.mat, aes(me.exp, me.conf)) +
  geom_tile(color="white",aes(fill = estimate)) +
  geom_text(aes(label = round(estimate, 2))) +
  scale_fill_gradient2(low="#D55E00",
                      mid="white",
                      high = "#56B4E9",
                      midpoint=ref) +

```

```

labs(x=paste("% of total variance of HbA1c due to measurement error"),
     y=paste("% of total variance of BMI due to measurement error")) +
coord_equal()+
scale_y_continuous(breaks=unique(tot.mat[,1]))+
scale_x_continuous(breaks=unique(tot.mat[,1]))+
theme(panel.background = element_rect(fill='white', colour='grey'),
      plot.title=element_text(hjust=0),
      axis.ticks=element_blank(),
      axis.title=element_text(size=12),
      axis.text=element_text(size=10),
      legend.title=element_text(size=12),
      legend.text=element_text(size=10))

```

