L 3: Interpretation, correlation and confounding

Cars model

```
> summary(stop.model)

Call:
lm(formula = dist ~ speed + I(speed^2), data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-28.720  -9.184  -3.188   4.628  45.152

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.47014   14.81716   0.167    0.868
speed        0.91329    2.03422   0.449    0.656
I(speed^2)   0.09996    0.06597   1.515    0.136

Residual standard error: 15.18 on 47 degrees of freedom
Multiple R-squared:  0.6673,Adjusted R-squared:  0.6532
F-statistic: 47.14 on 2 and 47 DF,  p-value: 5.852e-12
```

We have very high p-values for the model terms.

Does this mean that the we can accept the corresponding null hypotheses, and conclude that all the parameter values could really be 0?

Ans: Absolutely not!

The p-values are testing whether the corresponding coefficients could really be zero *given that the other terms remain in the model* (i.e. are nonzero).

If the estimators for the various predictors are not independent, then dropping one term (setting it to zero) will change the estimates of the other coefficients and hence their p-values.

Hence should only drop one term at a time, refitting after each drop.

In which situation is it ok to drop all terms with high p-values simulataneously?

Balanced data from designed experiments

Odor data: Designed experiment to determine the effects of column temperature, gas/liquid ratio and packing height in reducing the unpleasant odour of a chemical product.

```
> data(odor , package ='faraway')
> odor[1:10,]
   odor temp gas pack
1    66   -1  -1    0
2    39    1  -1    0
3    43   -1   1    0
4    49    1   1    0
5    58   -1   0   -1
6    17    1   0   -1
7    -5   -1   0    1
8   -40    1   0    1
9    65    0  -1   -1
10    7    0   1   -1
```

Now compare the following two models:

```
> fmod = lm( odor ~ temp + gas + pack , data = odor )
> coef (fmod)
(Intercept)        temp         gas        pack
     15.200     -12.125     -17.000     -21.375
> rmod = lm( odor ~ temp + gas , data = odor )
> coef (rmod)
(Intercept)        temp         gas
     15.200     -12.125     -17.000
```

Orthogonality

Notice how the common coefficients are the same in both models. For this data, there is no correlation between any pair of variables. This is the orthogonality property. We can check it:

```
> cov(odor[,-1])
          temp       gas      pack
temp 0.5714286 0.0000000 0.0000000
gas  0.0000000 0.5714286 0.0000000
pack 0.0000000 0.0000000 0.5714286
```

- The diagonal property of this matrix demonstrates the orthogonality.
- This means we can test for the significance of these predictors without reference to what other variables have been included.
- Unfortunately, orthogonality doesn't happen by chance — orthogonality of $\mathbf{X}$ is an important characteristic of a good design.
- Orthogonaly rarely happens in observational data (e.g. surveys, cohort studies, ...)

Confounding

- Lack of independence between estimators creates difficulties in the interpretation of estimates.
- Correlation between parameter estimators arises from correlation between the variables to which the parameters relate.
- It is then not possible to entirely separate out their effects on the response.

Example: FEV and smoking in children

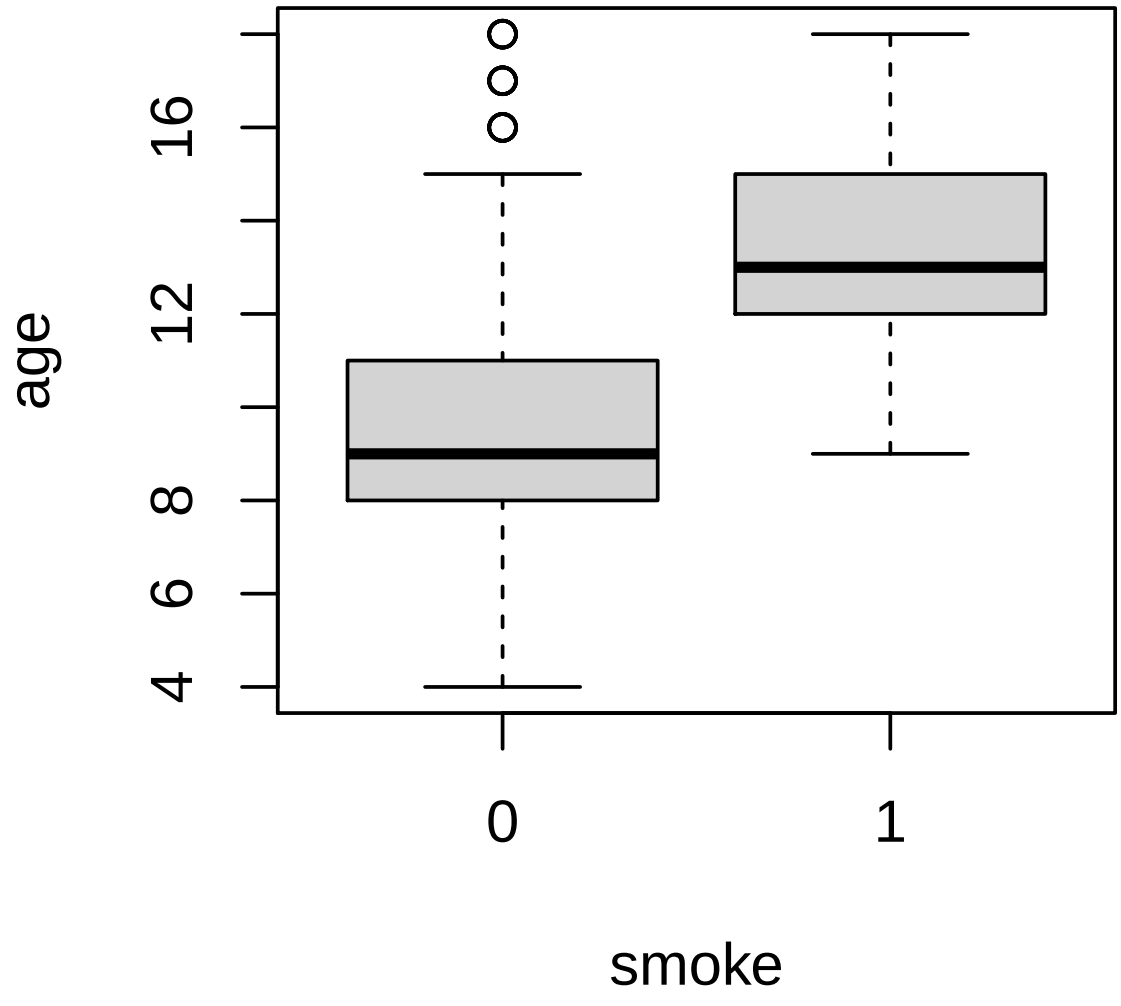
In adults, smoking reduces lung function as measured by FEV (lower FEV is bad). Researchers are interested in the effect of smoking on FEV in children. They collect data on 654 children aged 3-19.

```
> summary(lm(fev~smoke,data=fev))

Call:
lm(formula = fev ~ smoke, data = fev)
...

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.56614    0.03466  74.037  < 2e-16 ***
smoke        0.71072    0.10994   6.464 1.99e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

...
```

According to the results, smoking is beneficial! Average FEV is higher in smokers.

Why? One explanation is that older, male children are more likely to take up smoking and will have higher FEV by virtue of having larger lungs. We can adjust for age and sex using multiple linear regression.

FEV example: confounder variables age and sex



FEV example: adjust for confounder variables age and sex

```
> summary(lm(fev~smoke + age + male,data=fev))

Call:
lm(formula = fev ~ smoke + age + male, data = fev)

...

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.220614   0.080541   2.739  0.00633 **
smoke       -0.153922   0.077821  -1.978  0.04836 *
age          0.228610   0.007923  28.855  < 2e-16 ***
male         0.314244   0.042638   7.370  5.2e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

Blood pressure

We can show the effect of confounding through sampling. Consider the following code, which generates values for blood pressure entirely dependent on weight:

```
n <- 50; set.seed(7)
height <- rnorm(n,180,10)
weight <- height^2/400+rnorm(n)*5
bp <- 80 + weight/2 + rnorm(n)*10
```

Blood pressure

Now fit the linear model:

$$bp_i = \beta_0 + \beta_1 height_i + \beta_2 weight_i + \epsilon_i$$

```
> summary(lm(bp~height+weight))
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.76340 30.57422 1.072 0.2894
height 0.45497 0.26894 1.692 0.0973 .
weight 0.09462 0.27248 0.347 0.7299
```

Correlation between weight and height obscuring the true coefficient values (0 for height, 0.5 for weight).

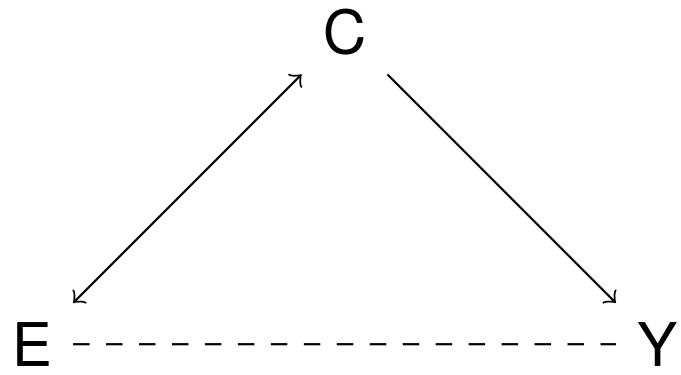
Confounding

Confounding occurs where there are additional variables (**confounders**) associated with both the response and the variable of interest.

The conditions for considering a factor *C* to be a potential confounder of the relationship between the explantory variable/exposure *E* and the response/disease *Y* are

1. C has a causal effect on Y
2. C is associated with E, but not a proxy/surrogate for E
3. C is not an intermediate step in the causal path between E and Y

We can describe the possible relationships between the response, explanatory variable of interest, and confounder using a causal diagram:



- where one sided arrows ($\rightarrow$) indicate causality,
- two sided arrow ($\leftrightarrow$) indicate association, and
- dashed lines ($- - -$) represent the relationship being studied.

Infering causality - Exercise

1. *Why is infering causality possible with data from designed experiments?*
2. *Why is infering causality difficult with data from observational studies?*

Infering causality

Is easy with data from designed experiments:

- In a balanced experimental design the variable of interest and the confounder variables are orthogonal.
- Random allocation of experimental units to variables controlled for (e.g. children to smoking!) breaks the relation between C and E and also means that C won't affect Y in a way that is biased according to E —in other words, an unobserved confounder just goes into the error.

Infering causality

Is difficult with data from observational studies:

- We can adjust for confounder variables in our model, but obviously we cannot adjust for hidden confounder variables.

Note: In contrast, if you only care about *prediction* and not causation, things are much simpler. Anything that associates with Y will improve your prediction even if the effect is not causal.

Example

- Say that we want to test the effect of two new drug treatments on blood pressure.
- What if we assigned all women to one treatment and all men to the other?
- It would be impossible to disentangle the effect of sex vs treatment.
- We could measure all possible confounders and include them in the model - how do we know we have them all?
- Instead randomly allocate patients to treatment. This breaks associations and means valid conclusions can be drawn.

Why is this important?

- In most cases, datasets are observational (except in medicine, where case control studies are widespread).
- Thus, you have to be very careful when drawing inference from fitted models.
- May be hidden confounders, may be impossible to separate out different effects in the model.
- Testing correlations between predictor variables, careful plots and using common sense checks (e.g. are there any obvious confounders?) can help.
- But it's important to highlight any limitations in the analysis that may have arisen from imperfect data.