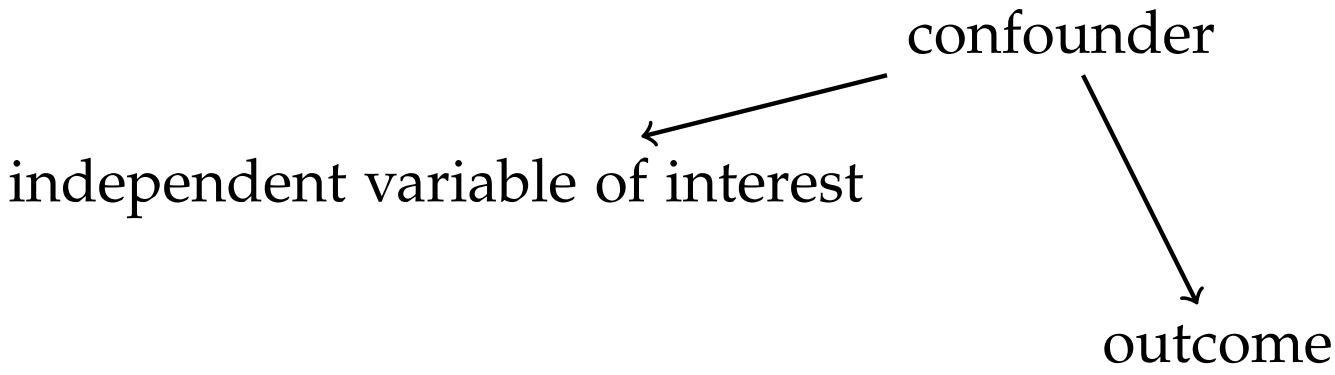
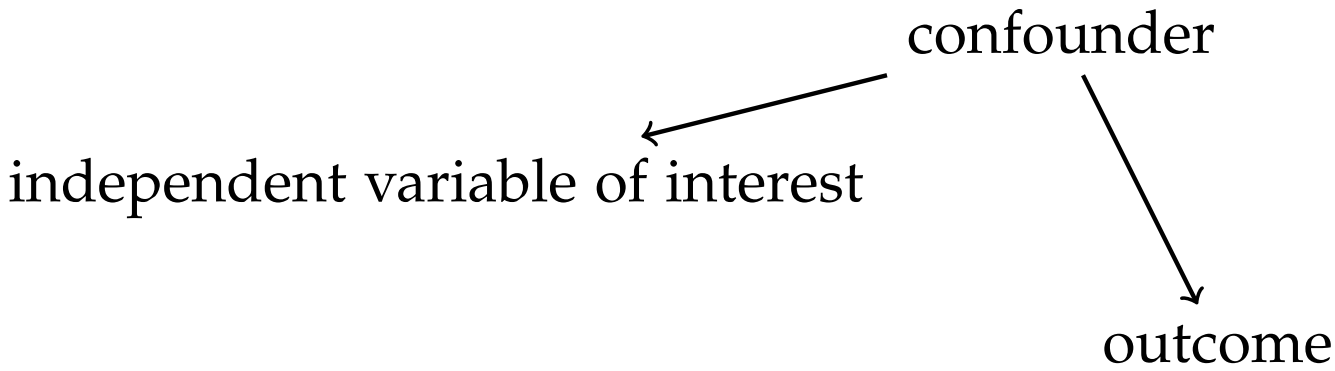
Logistic Regression: Confounding and Colinearity

Confounding in Logistic Regression



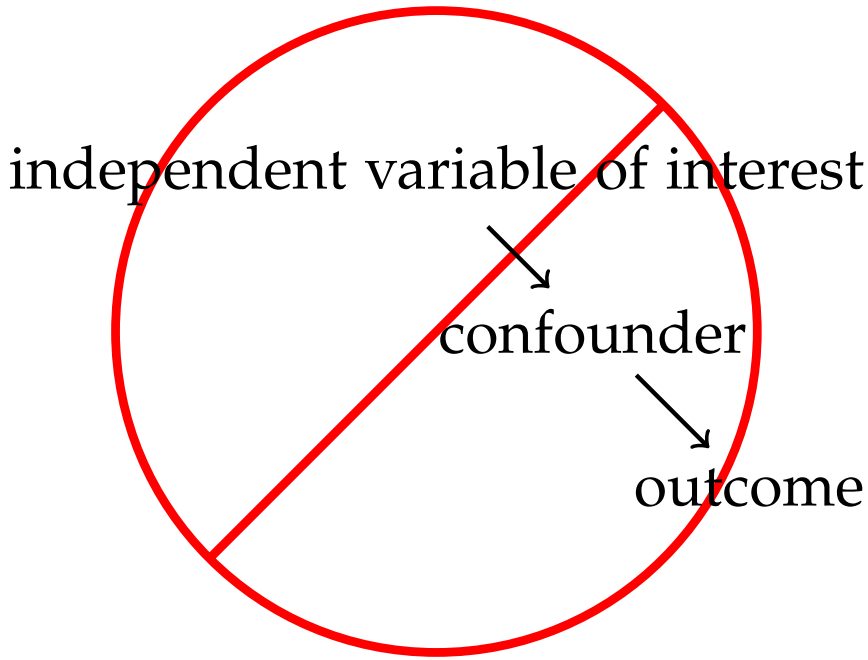
- This is the simplest situation
- A confounder must be a risk/protective factor for outcome
- A confounder must be associated with independent variable of interest

Confounding in Logistic Regression



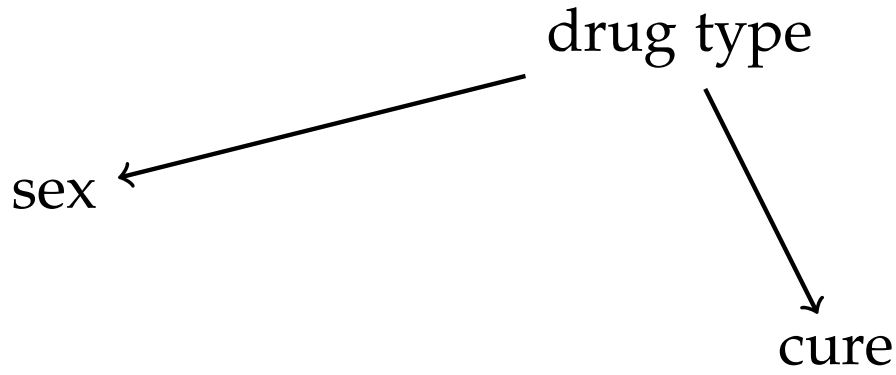
- All three variables are pairwise associated
- In a multivariate model with both independent variables included as predictors, the effect size of the variable of interest should be much smaller than the effect size of the variable of interest in the univariate model.

Confounding in Logistic Regression



- A confounder must not be an intermediate between the independent variable and the outcome.
- Unfortunately you cannot distinguish an intermediate from a confounder statistically
- Temporal ordering might help

Confounding: Drug Trial Toy Example



- Two types of drug: drug A and drug B.
- drug A cure rate is 30%, drug B cure rate is 50%
- no interaction between drug type and sex i.e drugs equally effective in both sexes.
- Due to bad study design all women randomised to drug A and all men randomised to drug B.

Drug Trial: Simulate in R

- Simulate trial of size 600 with 300 allocated to each drug type:

```
> drug <- as.factor(c(rep("A", 300), rep("B", 300)))
```

- Simulate perfect colinearity of drug and sex:

```
> sex  <- as.factor(c(rep("F", 300), rep("M", 300)))
```

- Simulate cure rates of 30% and 50%:

```
> cure <- c(rbinom(300, 1, 0.3), rbinom(300, 1, 0.5))
```

- Summarise the data:

```
> summary(cure.dat)
      cure       sex      drug
 Min.   :0.00   F:300   A:300
 1st Qu.:0.00   M:300   B:300
 Median :0.00
 Mean   :0.42
 3rd Qu.:1.00
 Max.   :1.00
```

Drug Trial: Fit Regression with `glm`

```
> output <- glm(cure ~ drug + sex, family = binomial)
> summary(output)
glm(formula = cure ~ drug + sex, family = binomial)

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8954     0.1272  -7.037 1.96e-12 ***
drugB         1.0961     0.1722   6.365 1.96e-10 ***
sexM              NA         NA      NA       NA
---
```

- sex automatically removed (as 2nd of 2 colinear variables)
- β for drug corresponds to odds ratio of $\exp(1.0961) = 2.99$
- Simulated odds of cure for drug A: $0.3/(1-0.3) = 0.428$ for drug B: $0.5/(1-0.5) = 1$. Odds ratio: $1.0/0.428 = 2.38$

Drug Trial: Fit Regression with `glm`

- Switch the order of predictors in input:

```
> output <- glm(cure ~ drug + sex, family = binomial)
> summary(output)
glm(formula = cure ~ drug + sex, family = binomial)

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8954     0.1272  -7.037 1.96e-12 ***
drugB         1.0961     0.1722   6.365 1.96e-10 ***
sexM              NA         NA      NA       NA
---
```

- drug removed (as 2nd of 2 colinear variables)
- same numerical results because of perfect colinearity

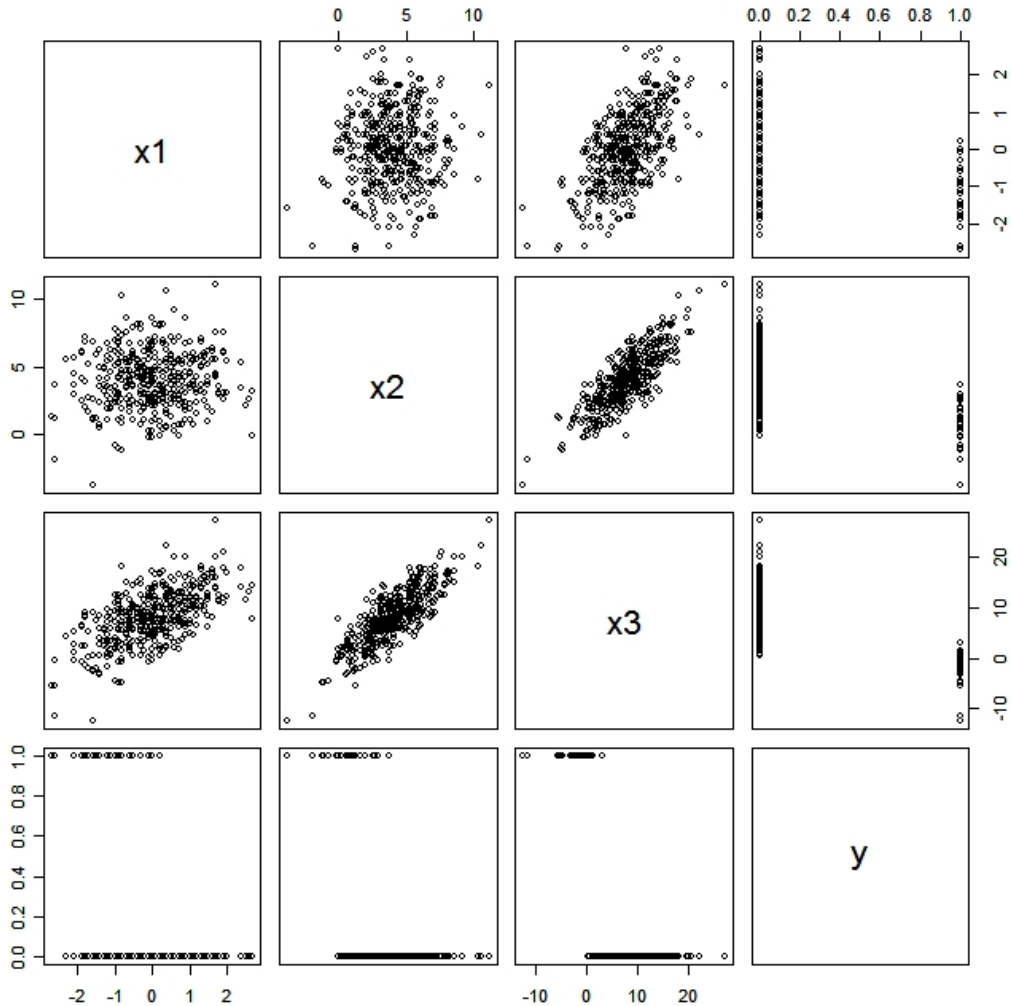
3-way Colinearity

- In this example 3 predictors are perfectly collinear.
- However, no 2 of variables are perfectly collinear.
- You can predict any of the variables perfectly from the other two.
- You can't predict any of the variables perfectly from just one of the others.

```
> x1 <- round(rnorm(400, mean=0, sd=1), 1)

> x2 <- round(rnorm(400, mean = 4, sd=2), 1)

> x3 <- 3*x1 + 2 *x2

> y <- rbinom(400, 1,
         exp(x1 + 2*x2 -3 * x3)/(1+ exp(x1 + 2*x2 -3 * x3)))
```

3-way Colinearity

- Pairwise correlation plot:



3-way colinearity: Fit Regression with `glm`

```
> output<-glm(y ~ x1+x2+x3, data = collinear.dat, family = binomial)
Warning message:
fitted probabilities numerically 0 or 1 occurred in:
glm.fit(x = X, y = Y, weights = weights, start = start, etastart = eta

> summary(output)

Call:
glm(formula = y ~ x1+x2+x3, family = binomial, data = collinear.dat)

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.7036     0.9001   1.893   0.0584 .
x1           -7.8162     1.8864  -4.144 3.42e-05 ***
x2           -4.6763     1.1203  -4.174 2.99e-05 ***
x3                NA         NA      NA       NA
---
```

- x3 has been eliminated, other variables reasonably estimated.

3-way colinearity: Meaning of Warnings

- This warning:

```
Warning message:
fitted probabilities numerically 0 or 1 occurred
```

  means that some of the within sample $\hat{\pi}_i$ are numerically one or zero (perfect classification).

- Means the effect is so strong that R cannot distinguish the predicted probabilities from 0 or 1.

- This warning:

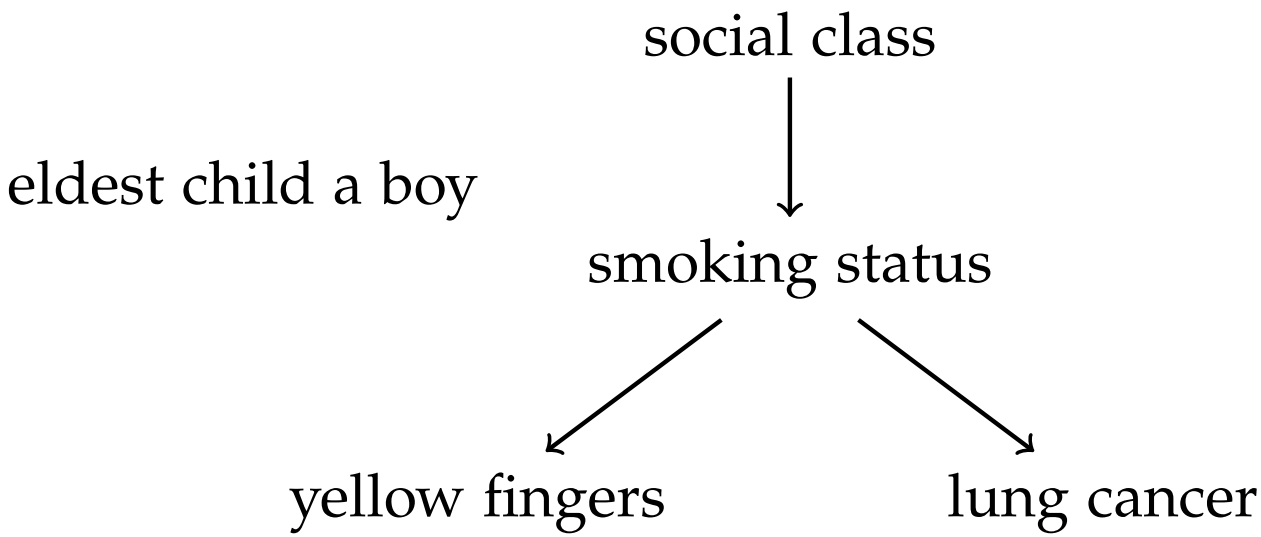
```
Coefficients:  (1 not defined because of singularities)
```

  is telling us to beware of perfect colinearity.

- In practice perfect colinearity, mostly occurs when a large number of categorical variables are used as predictors.

Identifying Confounding and Confounded Variables

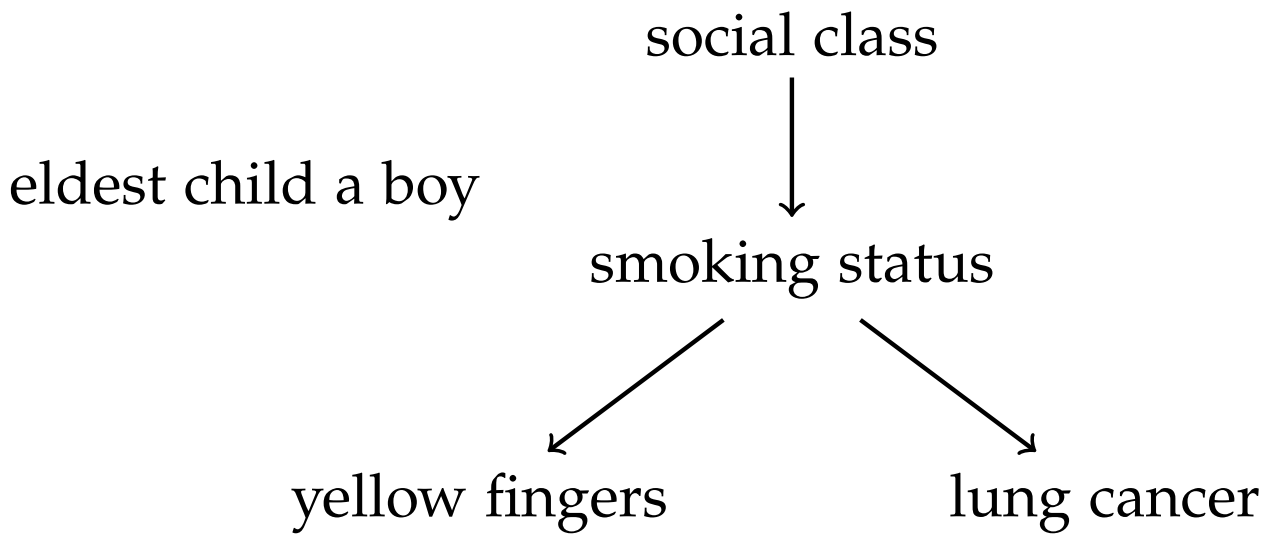
- Gold standard solution is to draw a causal diagram:



- Should aim to do "draw" something like this either on paper or at least in our heads.

Identifying Confounding and Confounded Variables

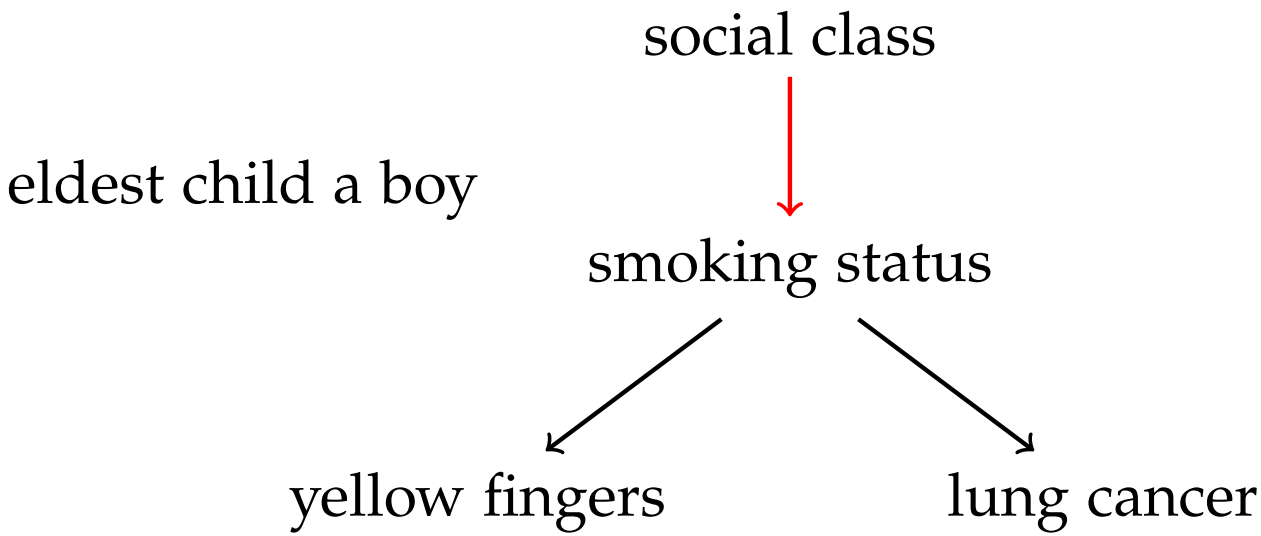
- Gold standard solution is to draw a causal diagram:



- Direction of arrow represents direction of causation
- Statistics can not tell us the direction of arrows

Identifying Confounding and Confounded Variables

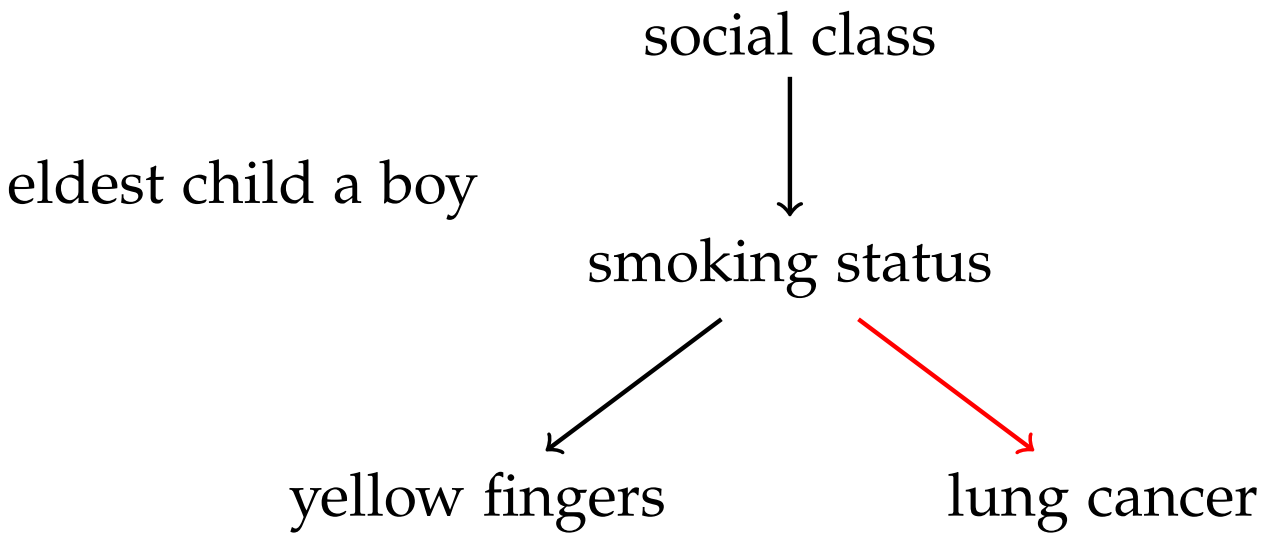
- Gold standard solution is to draw a causal diagram:



- Arrow direction can be determined by common sense and temporality
- e.g. common sense suggests smoking status does not *cause* a person to be more likely to have a manual job

Identifying Confounding and Confounded Variables

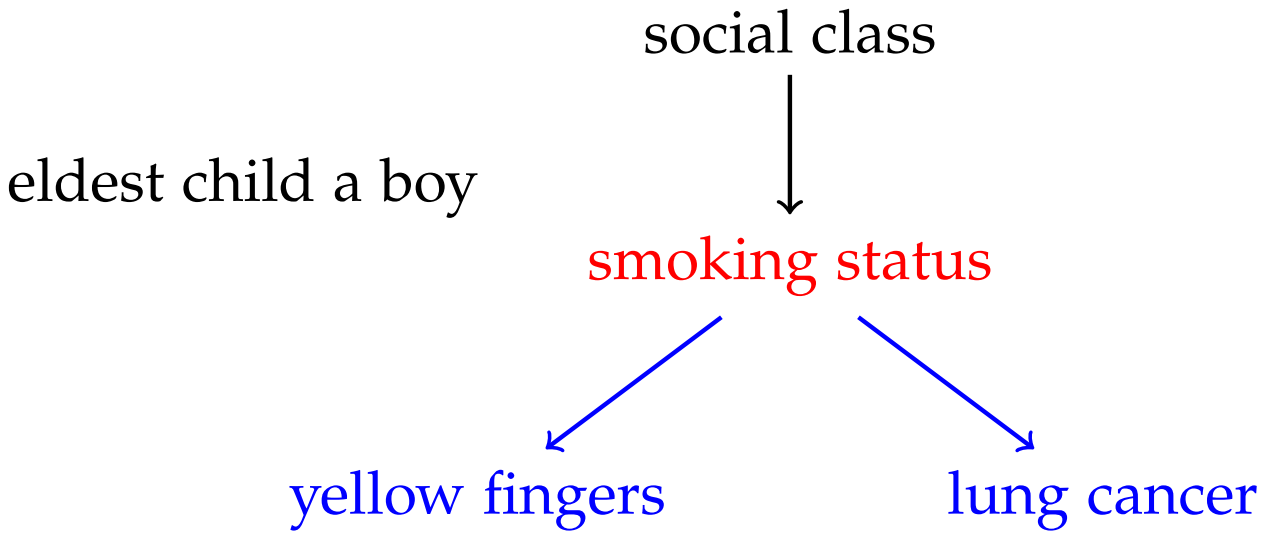
- Gold standard solution is to draw a causal diagram:



- Arrow direction can be determined by common sense and temporality
- e.g. temporality makes clear lung cancer is not a *cause* of a person being more likely to smoke.

Identifying Confounding and Confounded Variables

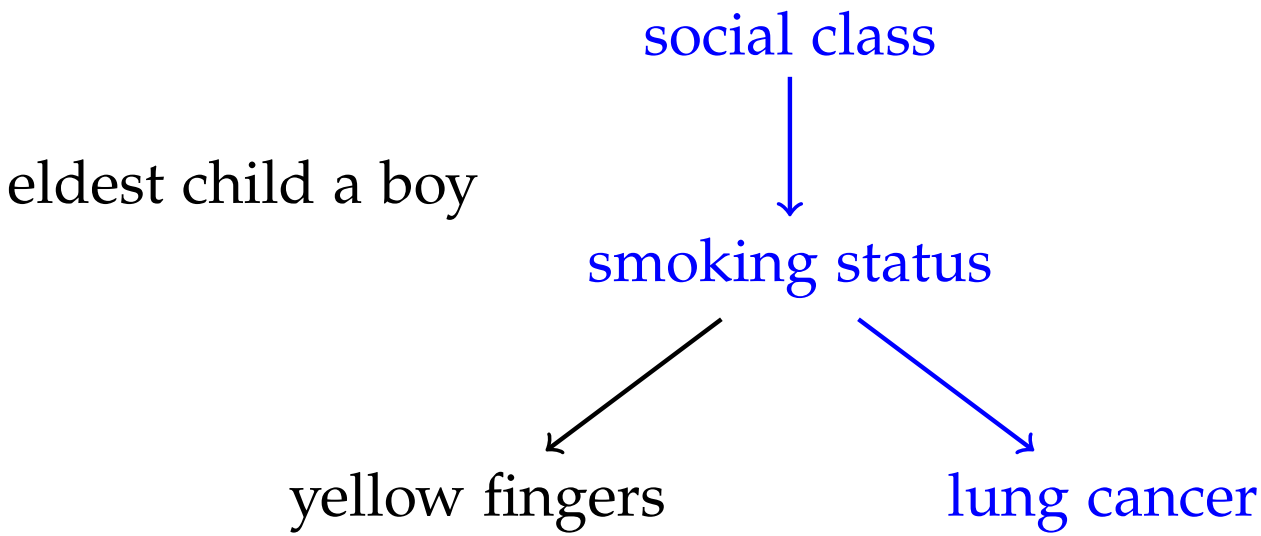
- Gold standard solution is to draw a causal diagram:



- Associations due to confounding can be identified by following arrows backwards to a common confounder
- e.g. association between yellow fingers and lung cancer is confounded by smoking status.

Identifying Confounding and Confounded Variables

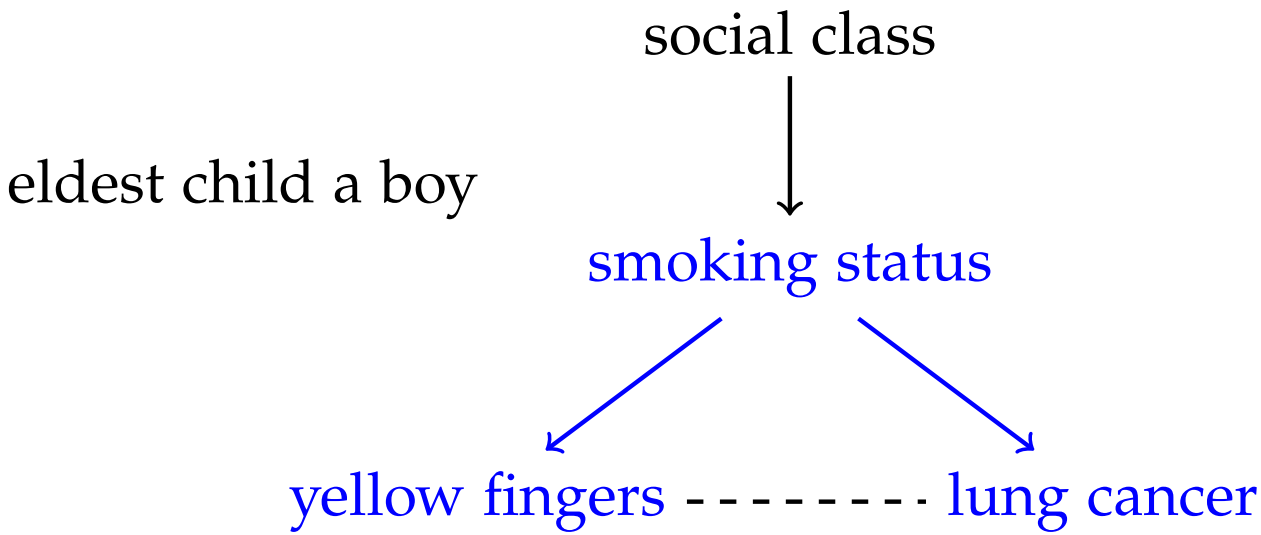
- Gold standard solution is to draw a causal diagram:



- Once you have a diagram in your mind you can "test it" using statistics.
- If you can follow the arrows from a variable to the outcome, you should see a univariate association

Identifying Confounding and Confounded Variables

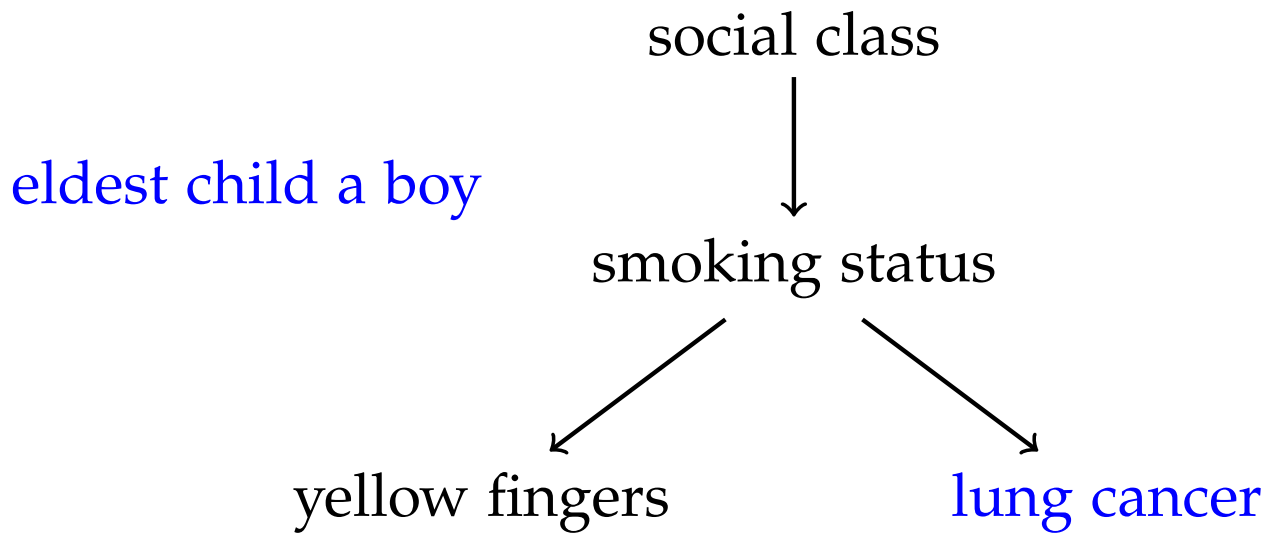
- Gold standard solution is to draw a causal diagram:



- Once you have a diagram in your mind you can "test it" using statistics.
- If you can follow the arrows backwards from a pair of variables to meet at a common variable you should see some evidence of a univariate association between them.

Identifying Confounding and Confounded Variables

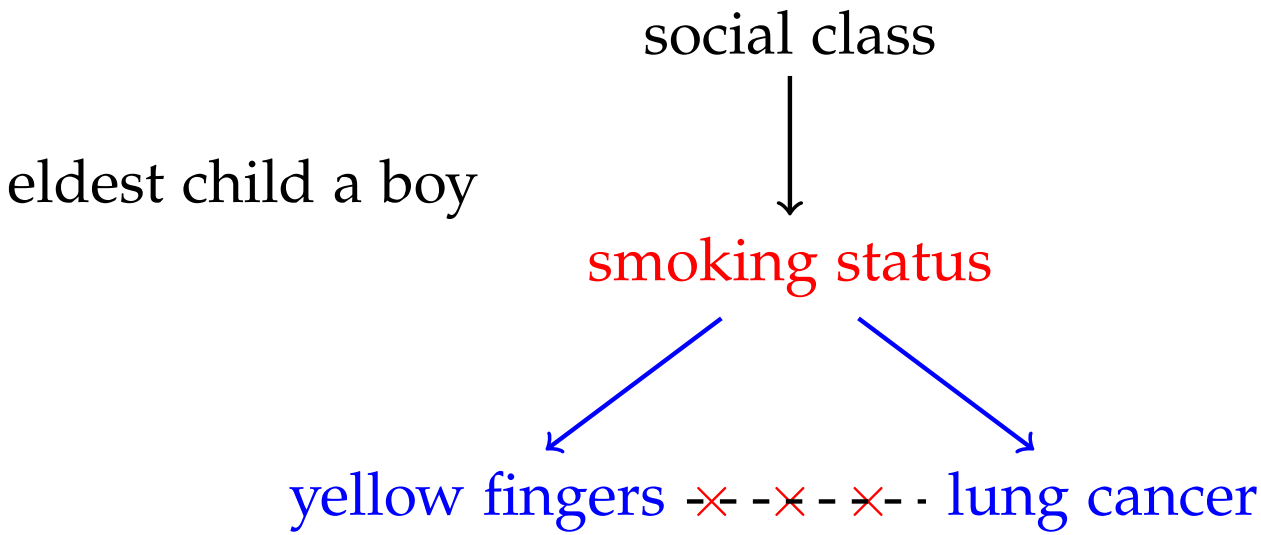
- Gold standard solution is to draw a causal diagram:



- Once you have a diagram in your mind you can "test it" using statistics.
- If you can't follow the arrows as described you shouldn't see a univariate association.

Identifying Confounding and Confounded Variables

- Gold standard solution is to draw a causal diagram:



- Once you have a diagram in your mind you can "test it" using statistics.
- If you fit a multivariate model including a "common ancestor" (confounder) and a predictor the univariate association with predictor should disappear or reduce.

Identifying Confounding and Confounded Variables

- Identifying confounding variables and confounded variables is an iterative process.

1. First try and draw a causal diagram based on common sense, temporality and univariate associations you observe.
2. Identify "common ancestor" variables as confounders.
3. Check that confounded associations reduce as you expect in a multivariate model.
4. If the causal diagram is inconsistent with the pattern of statistical association change it.

- **Give a list of confounded univariate associations and note which variables are the confounders for each confounded variable.**

Real Example: Low Birth Weight

- Low birth weight is of concern, because:
  1. Infant mortality rates and birth defect rates are very high for low birth weight babies.
  2. The Barker or "thrifty phenotype" hypothesis. Low birth weight babies are more susceptible to metabolic disorders later in life due to "fetal programming."
- Many factors (including diet, smoking habits, and receiving prenatal care) can affect the probability of carrying the baby to term and, consequently, the probability of delivering a baby of normal birth weight.

Low Birth Weight Example: Variables

Variable	R name
Low Birth Weight (0 = B.W. $\geqslant$ 2500g, 1 = B.W. $<$ 2500g)	low
Age of the Mother in Years	age
Weight in Pounds at the Last Menstrual Period	lwt
Race (1 = White, 2 = Black, 3 = Other)	race
Smoking Status During Pregnancy (1 = Yes, 0 = No)	smoke
History of Premature Labour (0 = None 1 = One, etc.)	ptl
History of Hypertension (1 = Yes, 0 = No)	ht
Presence of Uterine Irritability (1 = Yes, 0 = No)	ui
No. Physician Visits During the First Trimester (0,1,2,...)	ftv
Birth Weight in Grams	bwt

Low Birth Weight Example: Variable Summary

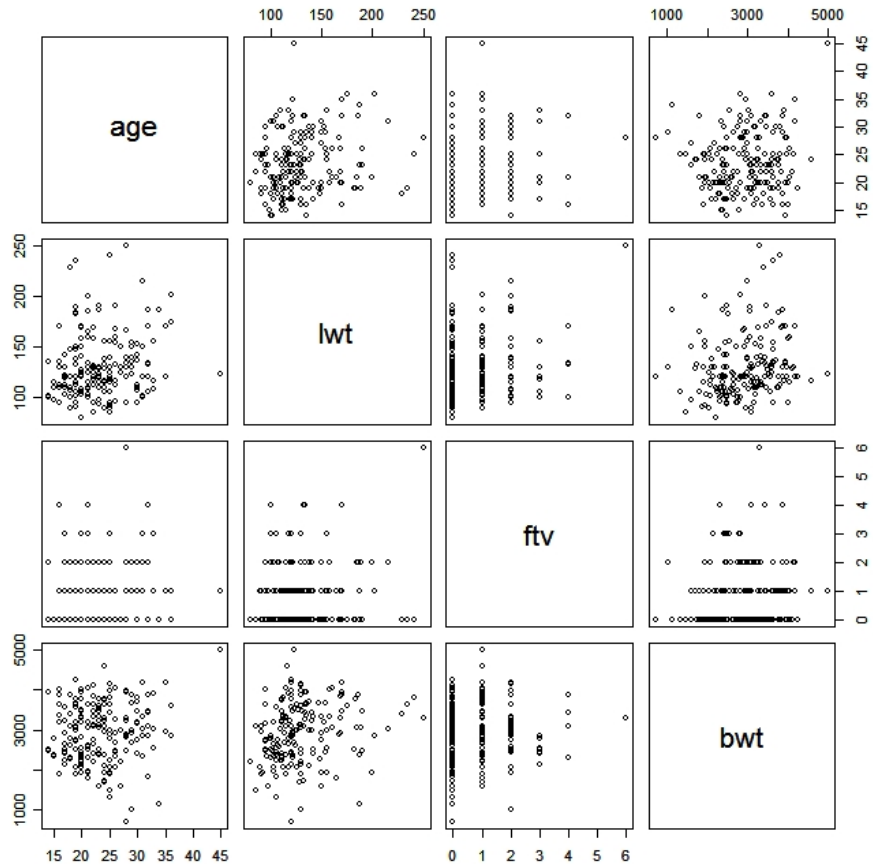
```
> summary(lbw.dat)

      low               age              lwt          race    smoke
 Min.   :0.0000   Min.   :14.00   Min.   : 80.0   1:96   0:115
 1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   2:26   1: 74
 Median :0.0000   Median :23.00   Median :121.0   3:67
 Mean   :0.3122   Mean   :23.24   Mean   :129.8
 3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0
 Max.   :1.0000   Max.   :45.00   Max.   :250.0

   ptl      ht          ui             ftv              bwt
 0:159    0:177    0:161    Min.   :0.0000   Min.   : 709
 1: 24    1: 12    1: 28    1st Qu.:0.0000   1st Qu.:2414
 2:  5                      Median :0.0000   Median :2977
 3:  1                      Mean   :0.7937   Mean   :2945
                            3rd Qu.:1.0000   3rd Qu.:3475
                            Max.   :6.0000   Max.   :4990
```

- low counts; reasonable to believe that history of at least one premature birth an important risk factor while precise no. premature births less important, so will collapse.

Low Birth Weight Example: Pairwise Correlations

- Some correlations e.g., age and lwt; nothing too extreme.

```
> pairs(list(age=lbw.dat$age, lwt=lbw.dat$lwt,
      ftv=lbw.dat$ftv, bwt=lbw.dat$bwt))
```



Low Birth Weight Example: `low` and `bwt`

Variable	R name
Low Birth Weight (0 = B.W. $\geqslant$ 2500g, 1 = B.W. $<$ 2500g)	low
Birth Weight in Grams	bwt

- Note that low (the outcome variable) is a dichotomised version of bwt which is a continuous variable.
- Because we want a dichotomous outcome variable for this example we will throw out bwt.
- In general this is not a good thing to do in a real data analysis as by dichotomising a variable you discard information.

Low Birth Weight Example: `low` regressed on `age`

```
> output <- glm(low ~ age, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.38458    0.73212   0.525    0.599
age         -0.05115    0.03151  -1.623    0.105

$OR
      age
0.9501333

$OR.ci
[1] 0.8932232 1.0106694
```

- No significant univariate association
- Mild evidence that risk of low $= 1$ goes down with age

Low Birth Weight Example: `low` regressed on `lwt`

```
> output <- glm(low ~ lwt, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99831    0.78529   1.271   0.2036
lwt         -0.01406    0.00617  -2.279   0.0227 *

$OR
    lwt
0.98604

$OR.ci
[1] 0.9741885 0.9980358
```

- Association is significant at 95% level.
- Evidence that $\mathbb{P}(\text{low} = 1)$ decreases as mothers weight at last menstrual cycle increases. (Clinically significant?)

Low Birth Weight Example: `low` regressed on `race`

```
> output <- glm(low ~ race, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.1550     0.2391  -4.830 1.36e-06 ***
race2         0.8448     0.4634   1.823   0.0683 .
race3         0.6362     0.3478   1.829   0.0674 .

$OR
   race2    race3
2.327536 1.889234

$OR.ci
          [,1]     [,2]
[1,] 0.9385074 5.772384
[2,] 0.9554579 3.735596
```

- Mild evidence that $\mathbb{P}(\text{low} = 1)$ is greater for blacks (race=2) than whites (race=1) and greater for other races (race=3) than whites.

Low Birth Weight Example: `low` regressed on `smoke`

```
> output <- glm(low ~ smoke, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0871     0.2147  -5.062 4.14e-07 ***
smoke1        0.7041     0.3196   2.203   0.0276 *
---

$OR
  smoke1
2.021944

$OR.ci
[1] 1.080660 3.783111
```

- Good evidence that $\mathbb{P}(\text{low} = 1)$ is greater amongst smokers than non-smokers

Low Birth Weight Example: `low` regressed on `ptl`

```
> output <- glm(low ~ ptl, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.0571     0.1813  -5.831  5.5e-09 ***
ptl1          1.4626     0.4144   3.529 0.000417 ***

$OR
    ptl1
4.317073

$OR.ci
[1] 1.916128 9.726449
```

- A history of premature labour increases the odds in favour of low = 1 by an estimated factor of 4.3.

Low Birth Weight Example: `low` regressed on `ht`

```
> output <- glm(low ~ ht, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8771     0.1650  -5.315 1.07e-07 ***
ht1           1.2135     0.6083   1.995   0.0461 *
---

$OR
     ht1
3.365385

$OR.ci
[1]  1.021427 11.088221
```

- A history of hypertension seems to increase $\mathbb{P}(\text{low} = 1)$.

Low Birth Weight Example: `low` regressed on `ui`

```
> output <- glm(low ~ ui, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9469     0.1756  -5.392 6.97e-08 ***
ui1           0.9469     0.4168   2.272   0.0231 *

$OR
     ui1
2.577778

$OR.ci
[1] 1.138905 5.834499
```

- Uterine irritability seems to increase $\mathbb{P}(\text{low} = 1)$.

Low Birth Weight Example: `low` regressed on `ftv`

```
> output <- glm(low ~ ftv, data = lbw.dat, family=binomial)
> logistic.regression.or.ci(output)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6868     0.1948  -3.525 0.000423 ***
ftv          -0.1351     0.1567  -0.862 0.388527

$OR
      ftv
0.8736112

$OR.ci
[1] 0.6425933 1.1876819
```

- Number of physician visits has no obvious effect on $\mathbb{P}(\mathsf{low} = 1)$.

Low Birth Weight Example: Multivariate Regression

```
Call:
glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
    ftv, family = binomial, data = lbw.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6305  -0.7894  -0.5094   0.9119   2.2257

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.644476   1.223889   0.527  0.59849
age         -0.039548   0.038305  -1.032  0.30186
lwt         -0.015078   0.007034  -2.143  0.03207 *
race2        1.218791   0.533168   2.286  0.02226 *
race3        0.819439   0.450466   1.819  0.06890 .
smoke1       0.859459   0.409836   2.097  0.03599 *
ptl1         1.218512   0.463015   2.632  0.00850 **
ht1          1.860429   0.708161   2.627  0.00861 **
ui1          0.719299   0.463419   1.552  0.12062
ftv          0.050900   0.175456   0.290  0.77174
---
```

Low Birth Weight Example: Compare Univariate/Multivariate Regression

	Multivariate		Univariate	
Variable	OR	CI	OR	CI
age	0.96	(0.89, 1.04)	0.95	(0.89, 1.01)
lwt	0.99	(0.97, 1.00)	0.99	(0.97, 1.00)
race2	3.39	(1.19, 9.62)	2.02	(0.93, 5.77)
race3	2.27	(0.94, 5.49)	1.89	(0.96, 3.74)
smoke1	2.36	(1.06, 5.27)	2.02	(1.08, 3.78)
ptl1	3.38	(1.36, 8.38)	4.32	(1.92, 9.73)
ht1	6.42	(1.60, 25.75)	3.37	(1.02, 11.09)
ui1	2.05	(0.83, 5.09)	2.58	(1.14, 5.83)
ftv	1.05	(0.75, 1.48)	0.87	(0.64, 1.19)

- Hard to draw any firm conclusions, as confidence intervals so wide.
- Possible ui1 was a confounded association: need to look at its correlation with other predictors.